Institut des ACTUAIRES

ENSAE IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 09/11/2022

Par : Yao Ge

Titre : Cause-of-death modelling and mortality risk

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Nom : Caroline Hillairet

Membres présents du jury de l'Institut des Actuaires

Caroline Hillairet

Guillaume Biessy

Etienne Kichy

Secrétariat:

Bibliothèque:

Entreprise : SCOR

Signature :

Directeur du mémoire en entreprise :

Nom : Keunoo Chang
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Abstract

Following an improvement for almost a century in the US, life expectancy at birth has recently edged down, due to the deterioration of mortality rates among certain ages. Indeed, there has been an increase of mortality in younger ages and a slowdown of mortality improvement in elder ages.

For (re-)insurance companies which hold exposure on life risks, mortality rates forecast impact their profitability as well as their solvency. Thus appropriate evaluation and precise understanding of mortality experience are essential. In addition to the analysis of the historical mortality rates on an aggregate basis, cause-of-death modelling could shed light on refined understanding of trends among different age groups and genders.

More granular studies are needed to further understand the source of the deterioration of mortality rates. This thesis attempts to use a cause-of-death view instead of an aggregate view to model mortality risk, and investigates the suitability of three cause-of-death modelling approaches under the framework of mortality risk modelling, evaluates their respective strengths and weaknesses in a practical context, and intends to generate future mortality scenarios from a cause-of-death view.

After a review of the classical mortality risk modelling approach and an introduction to the cause-of-death modelling in the first two chapters, this thesis first tests a model based on independent cause assumption in the third chapter and the fourth chapter intends to employ an alternative modelling approach - Compositional data analysis techniques to produce more coherent cause-specific projection and future scenarios of mortality risk.

Keywords: Mortality; Cause-of-Death; Forecast; Mortality risk; Lee-Carter; Compositional data analysis; Life expectancy

Résumé

Après une amélioration pendant près d'un siècle, l'espérance de vie aux États-Unis à la naissance a récemment diminué, en raison de la détérioration des taux de mortalité pour certaines tranches d'âges. En effet, on constate une augmentation de la mortalité chez les jeunes et un ralentissement de l'amélioration de la mortalité chez les personnes âgées.

Pour les compagnies d'assurance et de réassurance qui ont une exposition aux risques de la vie, les prévisions de taux de mortalité ont un impact sur leur rentabilité et leur solvabilité. Une évaluation appropriée et une compréhension précise de l'expérience de la mortalité sont donc essentielles. En plus de l'analyse des taux de mortalité historiques sur une base agrégée, la modélisation des causes de décès pourrait permettre de mieux comprendre les tendances parmi les différents groupes d'âge et sexes.

Des études plus granulaires sont nécessaires pour mieux comprendre la source de la détérioration des taux de mortalité. Ce mémoire utilise une vision par cause de décès plutôt qu'une vision agrégée pour modéliser le risque de mortalité, et étudie la pertinence de trois approches de modélisation par causes de décès dans le cadre de la modélisation du risque de mortalité, évalue leurs forces et faiblesses respectives et propose une méthode pour générer des scénarios futurs de mortalité.

Après une revue de l'approche classique de la modélisation du risque de mortalité et une introduction des causes de décès dans les deux premiers chapitres, ce mémoire teste d'abord un modèle basé sur l'hypothèse des causes indépendantes dans le troisième chapitre, puis le quatrième chapitre teste une approche de modélisation alternative - les techniques d'analyse des données de compositions permettant une projection de taux de mortalité par cause de décès plus cohérente et une génération de scénarios pour le risque de mortalité.

Mots clés : Mortalité ; Cause du décès ; Prévision ; Risque de mortalité ; Lee-Carter ; Analyse des données compositionnelles ; Espérance de vie

Executive summary

Mortality in the US

Following an improvement of life expectancy for almost a century in the US, it has been reported that life expectancy at birth has edged down between 2015 and 2017 according to the Center of Disease Control and Prevention (CDC). The recent public health crisis diminished the slight recovery afterwards, and life expectancy dropped down by almost one year from 2020 to 2021. Put aside Covid-19, the reduction is also explained by various causes of death.

Objectives

More granular studies are needed to further understand the source of the deterioration of mortality rates. This thesis attempts to use a cause-of-death view instead of an aggregate view to model mortality risk. The objectives for each model tested are to:

- Understand the theory and behaviour of each of the different alternatives tested
- Evaluate the advantages and limits, notably the additional value and insights they provide
- Generate future scenarios to produce prediction intervals in line with the different trends of each cause-of-death

Framework

Human mortality database (HMD) is the current widely used mortality database in mortality risk modelling. It contains general population aggregate deaths at each age and year.

There exist two mortality databases which provide general population deaths by cause-of-death: Human Cause-of-Death database (HCD) and the Center of Disease Control and Prevention Underlying Cause-of-Death database (CDC). HCD data source is from National Center for Health Statistics (NCHS), it offers the number of deaths by cause, age group and year and it has the advantage of stable cause-of-death classification. CDC data is directly extracted from death certificates in the US, CDC data provides death numbers of each cause-of-death by single-year age and year. It has more variable choices such as educational level which allows for studying different sub-populations. This thesis has retained the CDC data for the cause-of-death analysis, because it provides single-year age format data and genuine data directly from the death certificates.

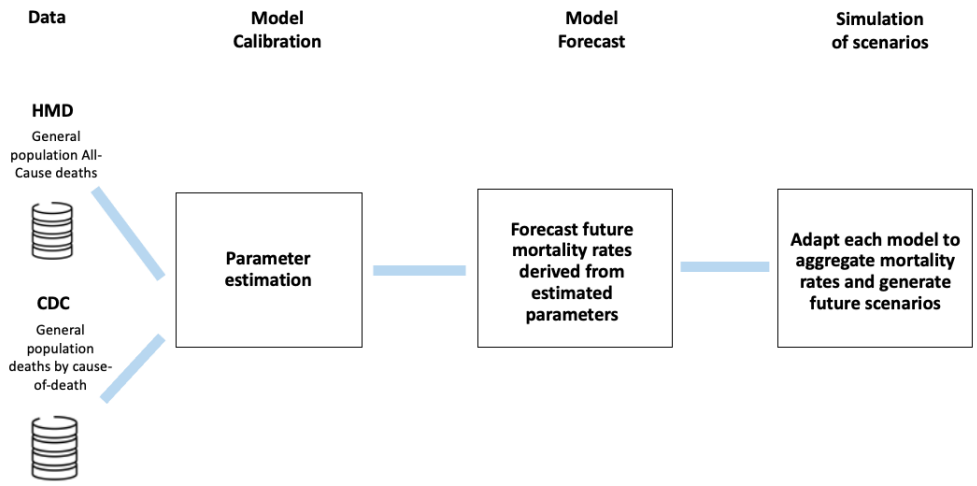


Figure 1: Modelling steps

Figure 1 presents the modelling framework followed in both All-Cause and By-Cause models. The All-Cause model refers to the classical Lee Carter model applied to the general population. All-Cause model is calibrated using HMD data and By-Cause models are calibrated on CDC. Models are calibrated from the data, used to forecast and generate scenarios.

Table 1 summarizes the three cause-of-death modelling approaches considered in this thesis. The first model tested emphasizes the individual dynamics laid in each cause which is modelled independently. The second and third model provide an alternative approach and tend to elaborate on the dependence between causes.

Table 1: Model comparison

Model name	**Modelling approach**	**Advantages**	**Limits**
Independent cause-specific model	Model each cause-of-death independently	Easy implementation and emphasize on individual cause characteristics	Unrealistic long-term forecast of aggregate mortality rates
CoDa Common Trend	- Pre-determine aggregate mortality rates forecast as constraint - Model each cause's proportion - Assume a common trend for every cause	- Coherence of cause-specific forecast with respect to aggregate mortality rates. - Explanatory ability - Risk transfer between cause-of-death and ages	- Long term forecast predicting dominance (over 70%) of *Drug-related* cause - Common Trend for each cause
CoDa Multi Trend	- Pre-determine aggregate mortality rates forecast as constraint - Model each cause's proportion - Assume an individual trend for every cause	- Coherence of cause-specific forecast with respect to aggregate mortality rates. - Explanatory ability - Specific trend evolution for each cause - Risk transfer between cause-of-death and ages	Long term forecast predicting dominance (over 70%) of *Drug-related* cause

Independent cause-specific model

The first modelling approach of this thesis is based on the main assumption that cause-specific central trajectories are independent. This assumption emphasizes the individual dynamics of each cause. Each cause's mortality rate is modelled by a classical Poisson log-bilinear model and the aggregate level mortality rate is expressed as the sum of cause-specific mortality rates. Prior to the modelling parts, historical cause-specific evolution

Figure 2: Aggregated and *Drug-related* log-mortality rates for male aged 20

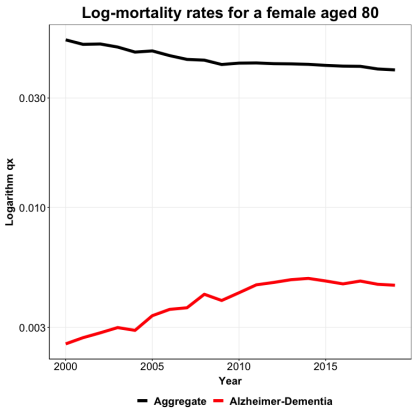


Figure 3: Aggregated and *Alzheimer-Dementia* log-mortality rates for female aged 80

and distribution among different ages are reviewed. Figure 2 and Figure 3 illustrate respectively the aggregate and *Drug-related* mortality rates evolution for a male aged 20, and mortality rates evolution for a female aged 80. It can be observed that the general population has a slightly declining trend while the specific causes show an opposite increasing trend.

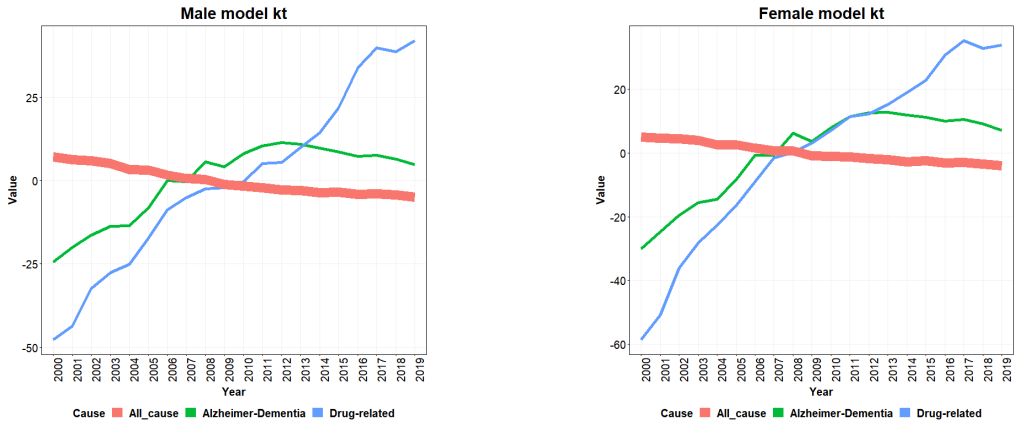


Figure 4: $\kappa_{t,i}$ for male and female

Figure 4 shows the trend parameter $\kappa_{t,i}$ of these two causes and All-cause model, which confirms the observation above: the aggregate mortality level deviation from the historical trend among the young and elder ages may be mainly conducted by the causes listed above.

The forecast is accomplished by extrapolation of the time index under the independent cause-specific assumption, therefore the historical trend of each cause is assumed to continue in the future. With 20 years of forecast, Figure 5 shows the residual life expectancy output by All-Cause and By-cause model. Life expectancy at birth for male and female will decline by two years in 20 years as per the model forecast, suggesting a

pessimistic forecast, even an unrealistic, especially regarding life expectancy at birth from this model.

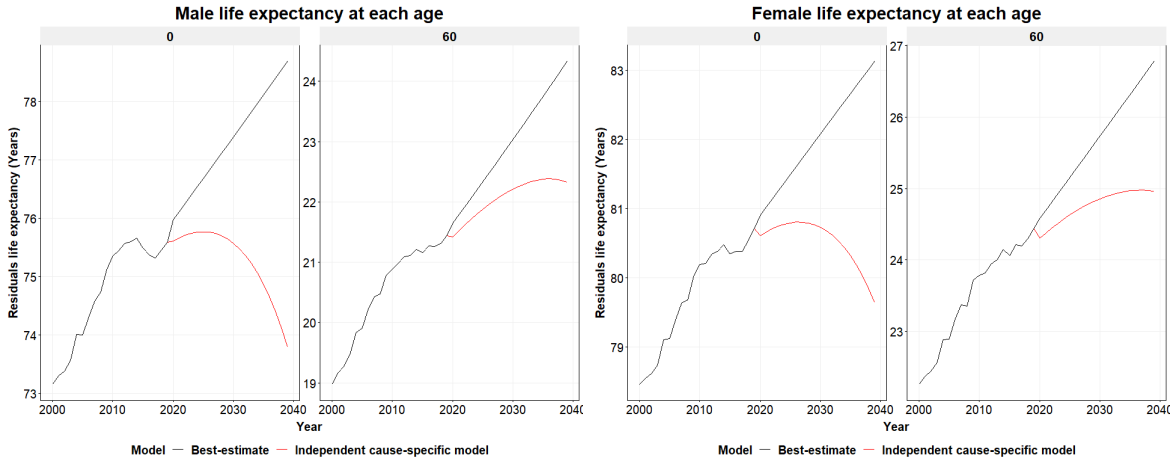


Figure 5: All-Cause and By-cause residual life expectancy at birth and 60

Figure 6 illustrates mortality rates forecast when the forecast horizon is extended to 60 years. In 60 years, if recent trend of each cause persists, mortality rates will at least double the classical All-Cause model forecast for all ages. This is mainly due to the linear extrapolation techniques used on each cause as it will follow independently its historical trend during the whole forecast horizon. Figure 7 shows the aggregate and *Drug-related* mortality rates forecast for a male at age 30, it could be seen that the cause *Drug-related* will become dominant in mortality rates, leading to unrealistic forecast. The underlying modelling approach could therefore be reviewed and contested by having ignored dependence between causes in the central trajectory.

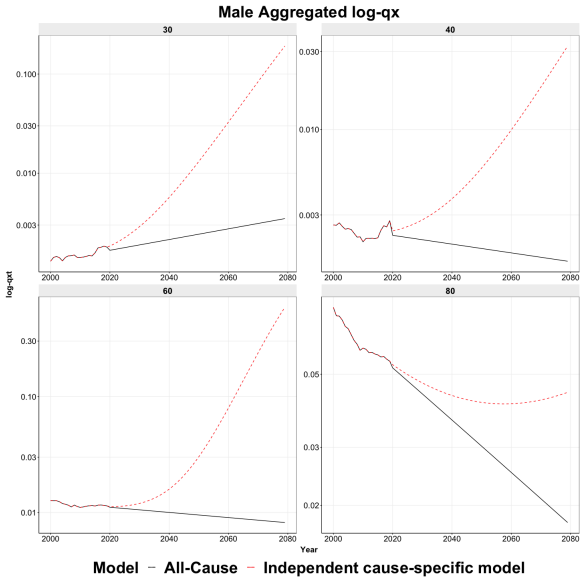


Figure 6: All-Cause and By-cause male mortality rates forecast for 60 years

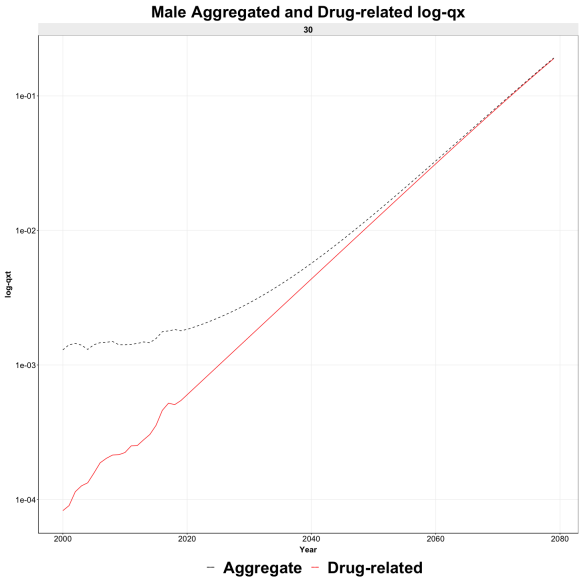


Figure 7: Aggregated and *Drug-related* log-mortality rates forecast for male aged 30

Compositional data analysis model

Given the unrealistic forecast of mortality rates resulting from the first model, additional constraints could be necessary to limit specific causes' trends, and ensure more coherent aggregate mortality rates. This chapter analyzes an alternative approach. Compositional Lee-Carter model is employed by shifting the target variable from absolute cause-specific mortality rates to the proportions of each cause in the future given historical evolution. A transformation (clr) is necessary to convert the proportions $d_{x,t,i}$ at age x, year t and cause i, which belong to a simplex, into numbers of a real space, allowing the application of broader statistical techniques. The Compositional Lee-Carter model could be expressed as below:

$$clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i}\kappa_t + \epsilon_{x,t,i} \tag{1}$$

Where the $\beta_{x,i}$ and κ_t have different interpretations with the classical All-Cause Lee-Carter model.

$\beta_{x,i}$ measures the age-and cause-specific sensitivity to trend factor κ_t, it describes the gain (or loss) of deaths for an age and a cause in relative terms. A positive $\beta_{x,i}$ for cause i associated with a positive κ_t means the cause i gains relatively more proportions compared to other causes.

An important mechanism and adding value of the CoDa framework is the concept of risk transfer, the reduction of the mortality risk in a cause or at an age will result in the increase in other causes or ages. The second part of this thesis investigates this alternative approach, the compositional Lee-Carter model could be further divided into two sub-models: Common Trend and Multi Trend with respect to the trend factor of cause. Common Trend model suggests a unique trend factor shared by each cause while Multi Trend allows for each cause a specific trend.

For these models, a constraint is defined on the aggregate mortality rates, which are predetermined by age and by year. Figure 8 illustrates the cause-specific forecast for a male at age 40, the forecast is more coherent with respect to aggregate mortality rates.

To generate scenarios from the central trajectories, the thesis proposes to:

- First, simulate different κ_t trajectories for the Common Trend model, and to simulate trajectories from a gaussian copula using the correlation matrix of trend residuals $\epsilon_{t,i}$ for the Multi Trend model.
- Calculate the resulting forecasted proportions for each cause-of-death.
- Calculate the resulting aggregate and by cause mortality rates.

This thesis notably proposes a method for the final step described above to convert the forecasted proportions by cause-of-death, age and year into absolute mortality rates, for a scenario different from the central trajectory. This is done by comparing the sum of transformed values in the inverse transformation operator C(exp(Y)) between a given scenario and the central trajectory.

$$C(exp(Y)) = [\frac{exp(Y_{x,t,i})}{\sum exp(Y_{x,t,i})}, ..., \frac{exp(Y_{x,t,j})}{\sum exp(Y_{x,t,i})}]$$

where Y represent the clr-transformed matrix.

An indicator I is defined as below $I_{x,t} = \frac{\sum exp(Y^{scenario}_{x,t,i})}{\sum exp(Y^{initial}_{x,t,i})}$ which represents the relative change of aggregate mortality rates for a male or female at age x and year t, which will be multiplied by the pre-determined aggregate mortality rates.

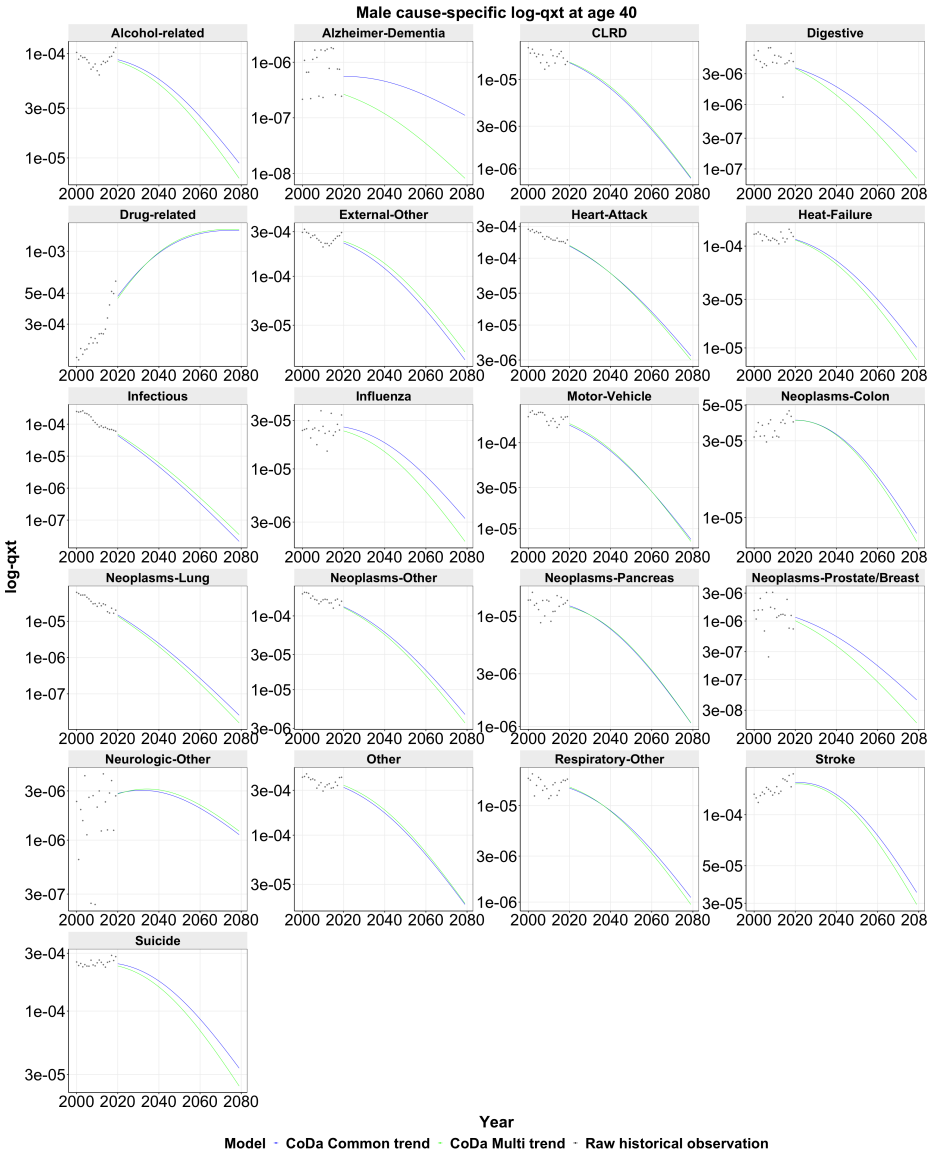


Figure 8: CoDa model male aged 40 mortality rates forecast for 60 years

Figure 9: Male cohort expected age at death interval

Then, the life expectancy deviations have been investigated. By looking at the expected age at death (attained age + life expectancy) at 0.5% and 99.5% levels presented in Figure 9, it can be observed that the results from All-Cause and CoDa models differ, the interval for CoDa Common Trend is notably narrower than the interval for the

All-Cause model. This could be partially explained by some lower resulting volatility of CoDa Common Trend's trend component.

As a last step, the scenario that corresponds to the 0.5% of expected age at death has been further analysed. Indeed, each forecasted scenario can be investigated on a cause-of-death granular level, providing an explanation of the cause proportion evolution, and insights to understand the sources of life expectancy decline.

Conclusion

This thesis has analysed three different cause-of-death modelling approaches: independent cause-specific; CoDa Common Trend and CoDa Multi Trend. CoDa models provide coherent mortality rates and interesting insights when analysing scenarios.

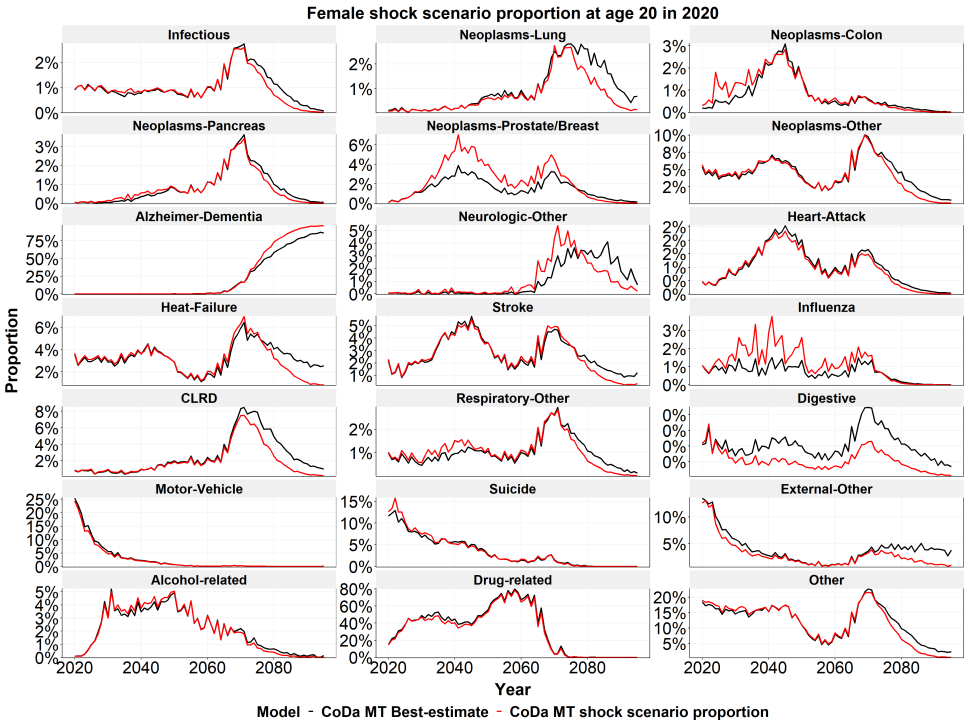


Figure 10: Best-estimate and Expected age at death at 0.5% level scenario of female cohort aged 20 in 2020

However, despite the more coherent forecast provided by CoDa models, it has been observed in the long-term forecast, certain causes such as *Drug-related* could still become dominant in proportions as displayed in Figure 10, *Drug-related* would occupy over 70% of forecasted mortality rates for the female cohort aged 20 in 2020 before 2070.

Therefore, the analysis should be further deepened by considering expert judgments, which could be applied to each cause-of-death to limit their evolution on a more rational scale. Some fundamental issues related to cause-of-death modelling remain to be solved as well, such as data quality following the change of classification standard.

In conclusion, the application of cause-of-death modelling is deemed to be premature without further adjustment and additional analysis for the mortality risk assessment. However, it can provide valuable insights of mortality trends and their evolution, detect the main drivers of aggregated mortality risk and explain extreme scenarios.

Note de synthèse

Mortalité aux États-Unis

Après une amélioration de l'espérance de vie pendant près d'un siècle aux États-Unis, on observe une baisse de l'espérance de vie à la naissance entre 2015 et 2017 selon le Center of Disease Control and Prevention (CDC). La récente amélioration de l'espérance de vie a été notamment entravée par la récente crise pandémique, ce qui a entrainé une chute de près d'un an en terme d'espérance de vie entre 2020 et 2021. En réalité, la récente détéoriation est également expliquée par d'autres causes de décès différentes du Covid-19.

Objectifs

Des études plus granulaires sont nécessaires pour mieux comprendre la source de la détérioration des taux de mortalité. Ce mémoire utilise une vision par cause de décès plutôt qu'une vision agrégée pour modéliser le risque de mortalité. Les objectifs de chaque modèle testé sont les suivants :

- Comprendre la théorie et le comportement de chacune des différentes alternatives testées.
- Evaluer les avantages et les limites, et notamment la valeur ajoutée et les perspectives qu'ils apportent.
- Générer des scénarios futurs pour produire des intervalles de projection en fonction des différentes tendances de chaque cause de décès.

Cadre de modélisation

La base de donnée utilisée classiquement pour la modélisation du risque de mortalité aux Etats Unis est la base Human Mortality Database (HMD). Elle contient les décès (toutes causes confondues) de la population générale pour chaque âge et chaque année.

Il existe deux bases de données de mortalité qui fournissent les décès de la population générale par cause de décès : Human Cause-of-Death Database (HCD) et la base de données du Center of Disease Control and Prevention (CDC). La base de donnée HCD provient de l'organisme National Center for Health Statistics (NCHS). Elle contient le nombre de décès par cause, par groupe d'âge et par année, et présente l'avantage d'une stabilité de la classification des causes de décès. D'autre part, la base CDC est issue directement des certificats de décès aux Etats Unis. Les données du CDC fournissent le nombre de décès pour chaque cause par âge et par année, elles proposent d'autres variables comme les niveaux d'éducation, permettant des analyses plus fines par type de population. La base de données CDC a été choisie pour cette étude car elles fournissent des données par âge, et des données authentiques à partir des certificats de décès.

La Figure 11 présente le cadre de modélisation suivi dans les modèles All-Cause et By-Cause. Le modèle All-Cause fait référence au modèle classique toutes causes de Lee

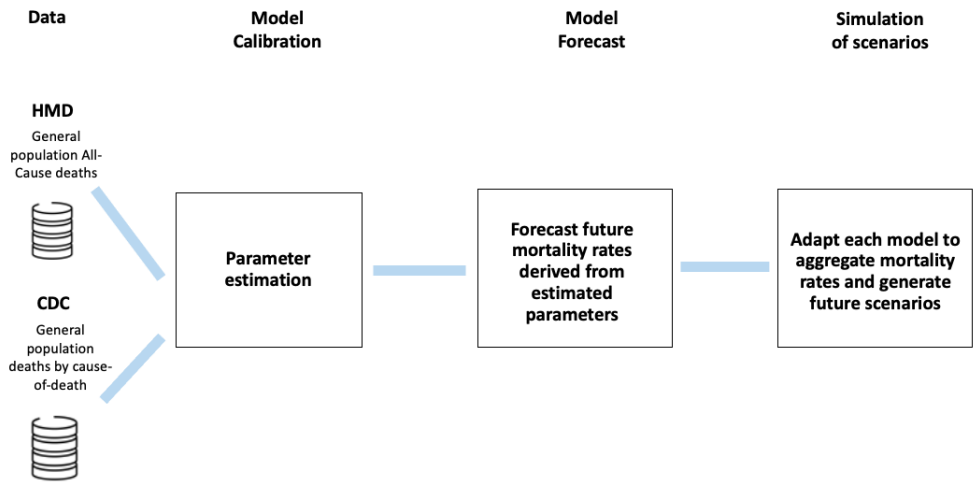


Figure 11: Les étapes de modélisation

Carter appliqué à la population générale. Le modèle All-Cause est calibré à l'aide de la base HMD alors que les modèles By-Cause sont calibrés via la base CDC. Une fois calibrés, les modèles sont utilisés pour projeter dans le futur et également générer des scénarios.

La Table 2 résume les trois approches de modélisation par cause de décès considérées dans ce mémoire. Le premier modèle testé met l'accent sur la dynamique individuelle de chaque cause qui est modélisée de manière indépendante. Les deuxième et troisième modèles fournissent une approche alternative et élaborent une dépendance entre les causes.

Table 2: Comparaison des modèles

Nom du modèle	**Approche de modélisation**	**Avantages**	**Limites**
Modèle avec hypothèse d'indépendance	Modélisation indépendante de chaque cause de décès	Mise en œuvre facile et accent mis sur les caractéristiques des causes individuelles	Projection à long terme irréaliste des taux de mortalité agrégés
CoDa Common Trend (tendance commune)	- Pré-détermination de la projection des taux de mortalité agrégés comme contrainte - Modélisation de la proportion de chaque cause - Hypothèse d'une tendance commune pour chaque cause	- Projection des taux de mortalité par cause en ligne avec le taux de mortalité agrégé - Capacité explicative - Transfert de risque entre la cause de décès et les âges	- Risque de projection long terme avec une domination (>70%) d'une cause spécifique (comme la drogue) - Tendance commune pour chaque cause
CoDa Multi Trend (tendances multiples)	- Pré-détermination des taux de mortalité agrégés prévus comme contrainte - Modélisation de la proportion de chaque cause - Hypothèse d'une tendance individuelle pour chaque cause	- Projection des taux de mortalité par cause en ligne avec le taux de mortalité agrégé - Capacité explicative - Evolution de la tendance spécifique à chaque cause - Transfert de risque entre la cause de décès et les âges	Risque de projection long terme avec une domination (>70%) d'une cause spécifique (comme la drogue)

Modèle avec hypothèse d'indépendance

La première approche de modélisation de ce mémoire est basée sur l'hypothèse principale que les trajectoires centrales de chaque cause sont indépendantes. Le taux de mortalité de chaque cause est modélisé par un modèle classique de Poisson log-bilinéaire et le taux de mortalité au niveau agrégé est exprimé comme la somme de ces taux de mortalité.

L'évolution historique est d'abord analysée. La Figure 12 et la Figure 13 illustrent respectivement l'évolution des taux de mortalité agrégés, les taux de la cause *Drug-related* pour un homme de 20 ans et *Alzheimer-Dementia* pour une femme de 80 ans. On observe que ces causes spécifiques présentent une tendance à la hausse alors que la population générale présente une légère tendance à la baisse.

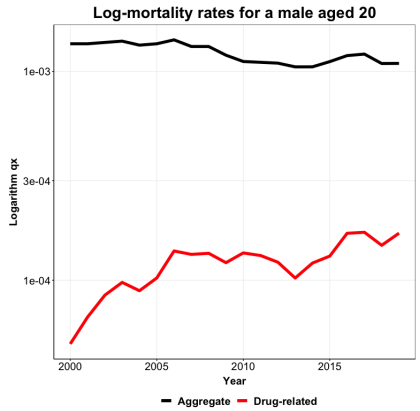


Figure 12: Taux de mortalité agrégés et de *Drug-related* pour un homme à l'âge 20

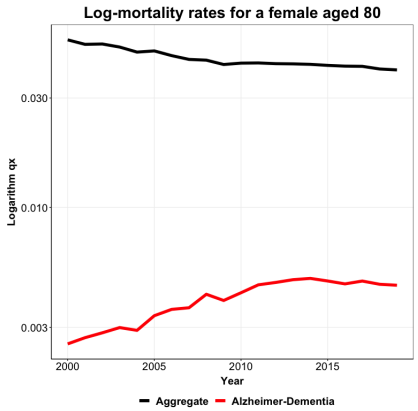


Figure 13: Taux de mortalité agrégés et de *Alzheimer-Dementia* pour une femme à l'âge 80

La Figure 14 montre le paramètre de tendance $\kappa_{t,i}$ du modèle Poisson log-bilinéaire, pour ces deux causes et celui du modèle toutes causes. Les fortes pentes observées pour ces deux causes pourraient expliquer les tendances haussières dans les graphes ci-dessus.

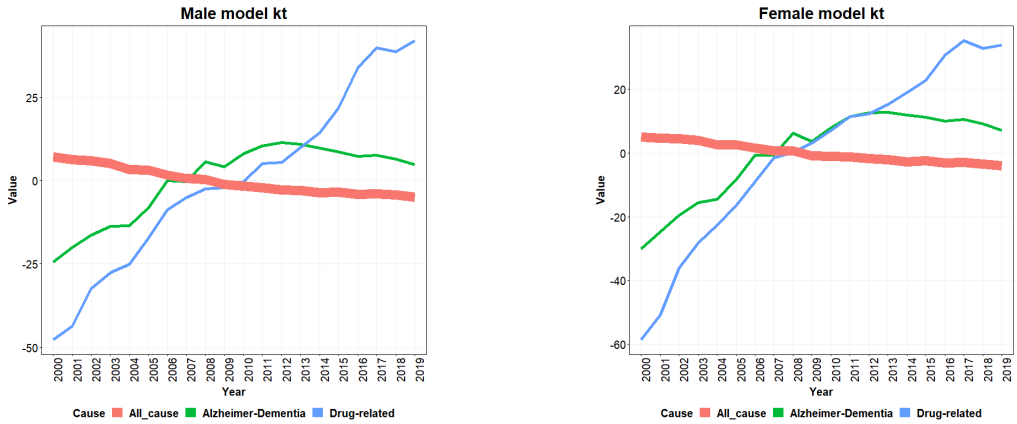


Figure 14: $\kappa_{t,i}$ de l'homme et de la femme

Dans le cadre de l'hypothèse d'indépendance des causes, la tendance historique de chaque cause est supposée se poursuivre selon une extrapolation de l'indice temporel. Avec 20 ans de projection, la Figure 15 montre les espérances de vie résiduelles pour le modèle toutes causes et par cause. L'espérance de vie à la naissance pour les hommes et les femmes diminueraient de deux ans dans 20 ans selon la projection du modèle, ce qui suggère une projection pessimiste pour l'espérance de vie à 60 ans, même irréaliste pour l'espérance de vie à la naissance.

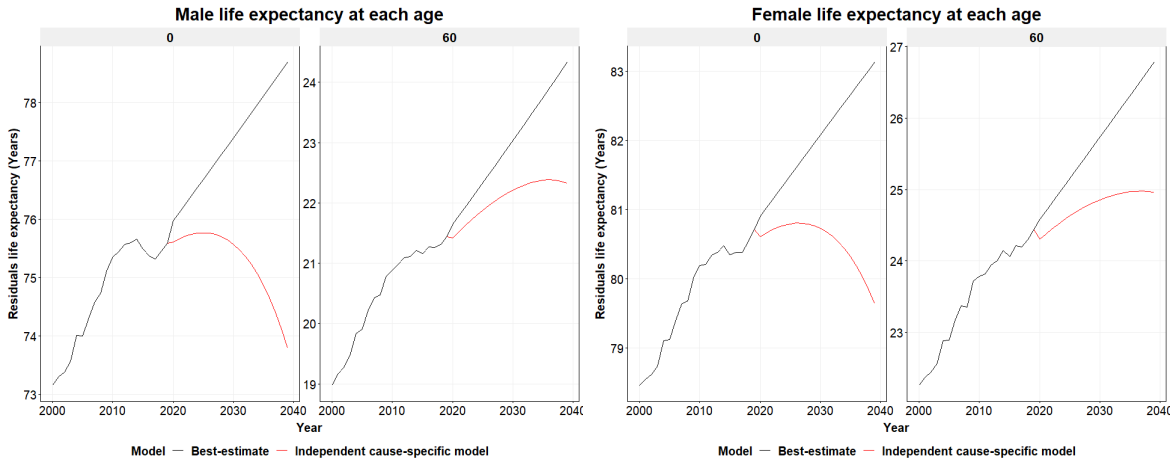


Figure 15: L'Espérance de vie résiduelle à la naissance et à l'âge 60 du modèle toutes causes et par cause.

La Figure 16 illustre les taux de mortalité projetés lorsque l'horizon de projection est étendu à 60 ans. Dans 60 ans, si la tendance observée actuellement pour chaque cause persiste, les taux de mortalité devraient au moins doubler par rapport aux projections du modèle classique pour tous les âges. Ceci est dû en premier lieu aux techniques d'extrapolation linéaire utilisées pour chaque cause, chaque cause suivant indépendamment sa tendance historique durant tout l'horizon de projection. La Figure 17 montre les projections de taux de mortalité agrégés et *Drug-related* pour un homme âgé de 30 ans, la cause *Drug-related* deviendrait dominante en absolu, entraînant des projections de taux de mortalité irréalistes. Suite à cette analyse, l'hypothèse d'indépendance utilisée par ce premier modèle peut donc être contestée.

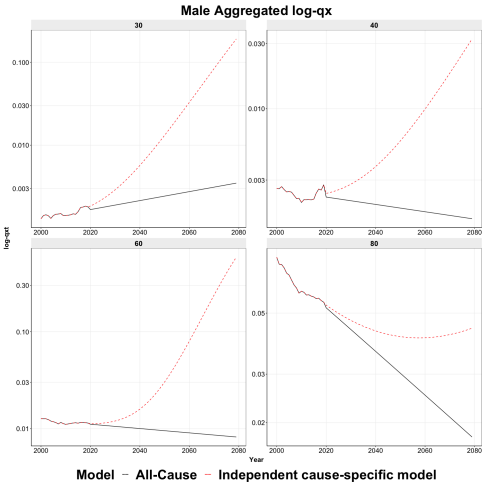


Figure 16: Projection à 60 ans du modèle toutes causes et par cause

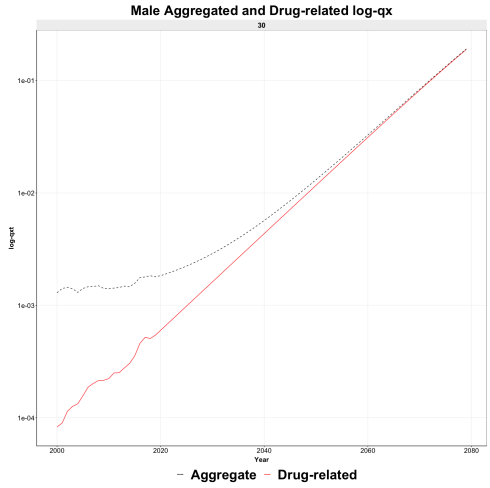


Figure 17: Projection du taux de mortalité agrégé et de *Drug-related* pour un homme de 30 ans

Modèle d'analyse des données de compositions

Compte tenu de la projection irréaliste des taux de mortalité résultant du premier modèle, certaines contraintes supplémentaires pourraient être nécessaires pour limiter les tendances des causes spécifiques et assurer des taux de mortalité agrégés plus cohérents.

Ce chapitre analyse une approche alternative. Le modèle de Lee-Carter de compositions est utilisé en changeant la variable cible des taux de mortalité absolus par cause en des proportions par cause compte tenu de l'évolution historique.

Une transformation (clr) est nécessaire pour convertir les proportions $d_{x,t,i}$ à l'âge x, l'année t et la cause i, permettant de passer d'un espace de simplex vers un espace réel et l'application de techniques statistiques. Le modèle de composition de Lee-Carter pourrait être exprimé comme suit :

$$clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i}\kappa_t + \epsilon_{x,t,i} \tag{2}$$

Où les $\beta_{x,i}$ et κ_t ont des interprétations différentes avec le modèle classique de Lee-Carter.

$\beta_{x,i}$ mesure la sensibilité par âge et par cause au facteur de tendance κ_t, il décrit le gain (ou la perte) de décès pour un âge et une cause en termes relatifs. Un $\beta_{x,i}$ positif pour la cause i associé à un κ_t positif signifie que la cause i gagne relativement plus de proportions par rapport aux autres causes.

Un mécanisme important d'un modèle de compositions (CoDa) est le concept de transfert de risque, la réduction du risque de mortalité dans une cause ou à un âge donné entraîne une augmentation dans d'autres causes ou à d'autres âges.

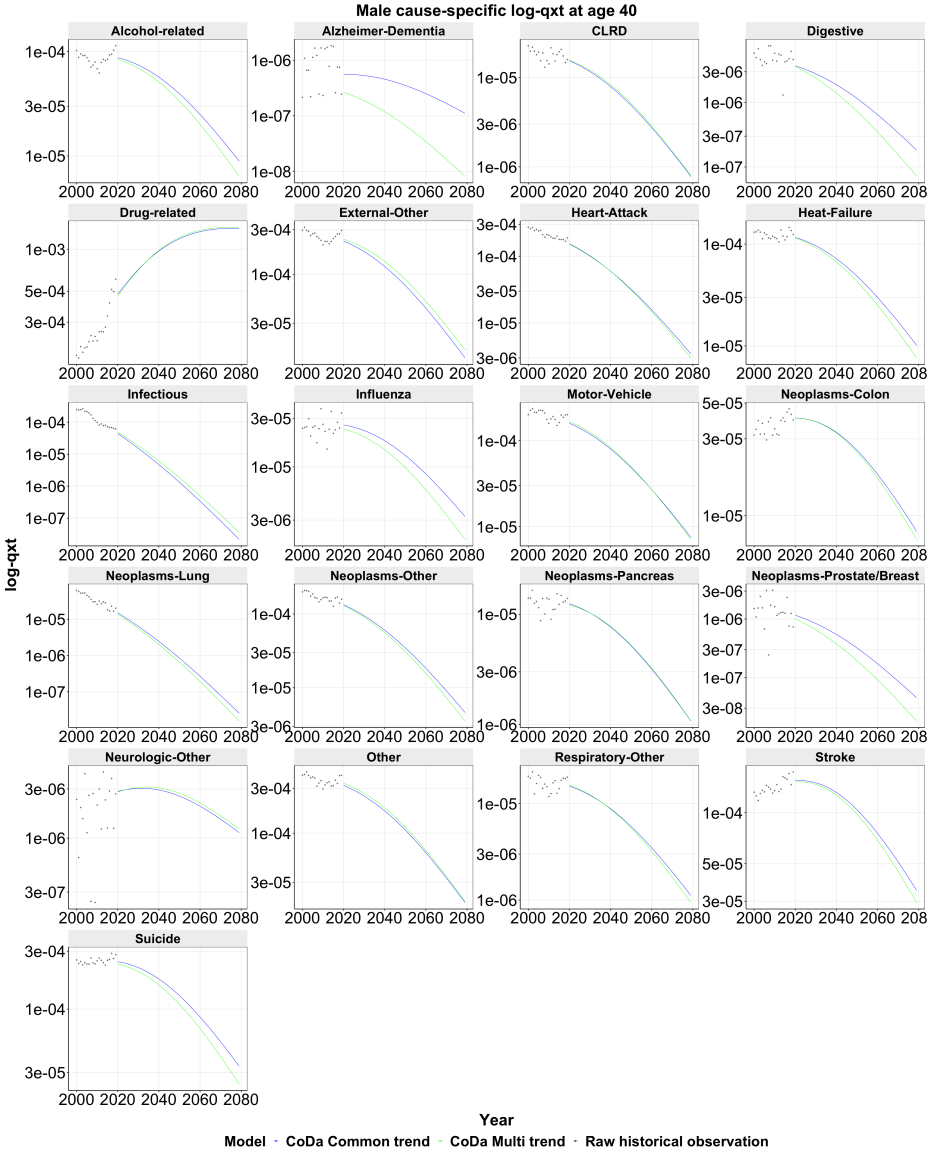


Figure 18: Projection par cause de décès à 60 ans des modèles de CoDa pour un homme de 40 ans

La deuxième partie de ce mémoire étudie cette approche alternative: le modèle Lee-Carter de compositions. Deux sous-modèles sont définis: Common Trend (CT ou tendance commune) et Multi Trend (MT ou tendance multiple). Le modèle Common Trend suggère un facteur de tendance unique partagé par chaque cause tandis que le modèle Multi Trend permet à chaque cause d'avoir une tendance spécifique.

Pour ces modèles, une contrainte est définie sur les taux de mortalité agrégés qui sont prédéterminés par âge et par année. La Figure 18 illustre les projections par cause de décès pour un homme de 40 ans, on observe que les projections sont plus cohérentes en ligne avec les taux de mortalité agrégés.

Pour générer des scénarios à partir des trajectoires centrales, ce mémoire propose de:

- Tout d'abord, simuler différentes trajectoires κ_t pour le modèle Common Trend et de simuler des trajectoires à partir d'une copule gaussienne en utilisant la matrice de corrélation des résidus de tendance $\epsilon_{t,i}$ pour le modèle Multi Trend
- Calculer les projections des proportions futures pour chaque cause de décès
- Calculez les taux de mortalité agrégés et par cause qui en résultent

Ce mémoire propose notamment une méthode pour l'étape finale décrite ci-dessus afin de convertir les proportions futures par cause de décès en taux de mortalités pour un scénario différent de la trajectoire centrale. Cette conversion s'effectue en comparant la somme des valeurs transformées dans l'opérateur de transformation inverse C(exp(Y)) entre un scénario donné et la trajectoire centrale.

$$C(exp(Y)) = [\frac{exp(Y_{x,t,i})}{\sum exp(Y_{x,t,i})}, ..., \frac{exp(Y_{x,t,j})}{\sum exp(Y_{x,t,i})}]$$

où Y représente la matrice des valeurs transformées.

Un indicateur I est défini comme suit : $I_{x,t} = \frac{\sum exp(Y^{scenario}_{x,t,i})}{\sum exp(Y^{initial}_{x,t,i})}$ qui représente la variation relative des taux de mortalité agrégés en pourcentage et qui sera multiplié par les taux de mortalité agrégés prédéterminés.

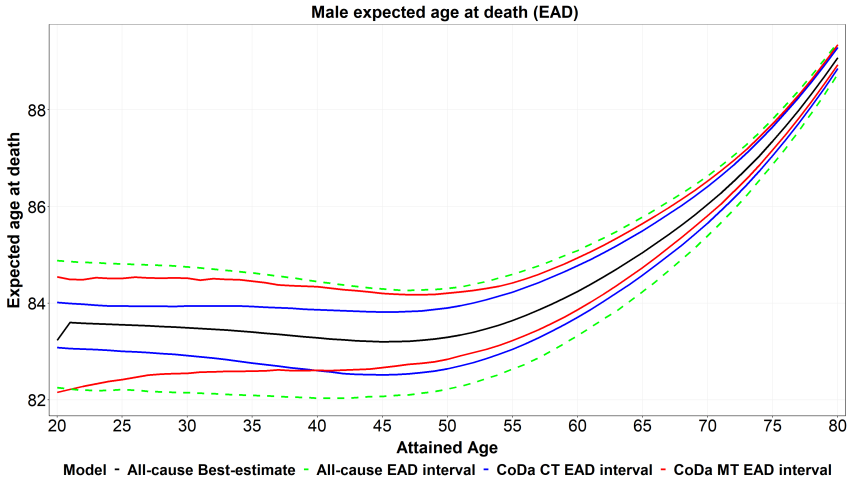


Figure 19: L'interval de l'âge attendu du mort du cohort pour l'homme

Ensuite, les écarts d'espérance de vie ont été étudiés. En examinant l'âge attendu au décès (âge atteint + espérance de vie) à des niveaux de 0,5% et 99,5% présentés dans la Figure 19, on peut observer que les résultats des modèles toutes causes et ceux des CoDa diffèrent, l'intervalle pour le modèle toute cause est plus étroit que celui du CoDa Common Trend. Cela pourrait s'expliquer en partie par une volatilité résultante plus faible de la composante de tendance κ_t issu du modèle de CoDa Common Trend par rapport à celui du modèle toutes causes.

Chaque scénario simulé peut enfin être analysé au niveau granulaire des causes de décès, ce qui permet d'expliquer l'évolution de la proportion des causes et de comprendre les sources du déclin de l'espérance de vie.

Conclusion

Ce mémoire a analysé trois approches différentes de modélisation des causes de décès: Modèle avec hypothèse d'indépendance, CoDa Common Trend (tendance commune), CoDa Multi Trend (tendance multiple). Les modèles de CoDa fournissent des taux de mortalité cohérents et des explications intéressantes lors de l'analyse des scénarios.

Cependant, malgré les projections plus cohérentes fournies par les modèles de CoDa, on observe que, dans les projections à horizon long terme, certaines causes telles que *Drug-related* pourraient devenir dominantes en termes de proportions, comme le montre la Figure 20 où la cause Drug related occuperait plus de 70% des taux de mortalités pour la cohorte femme âgée de 20 ans en 2020 jusqu'en 2070.

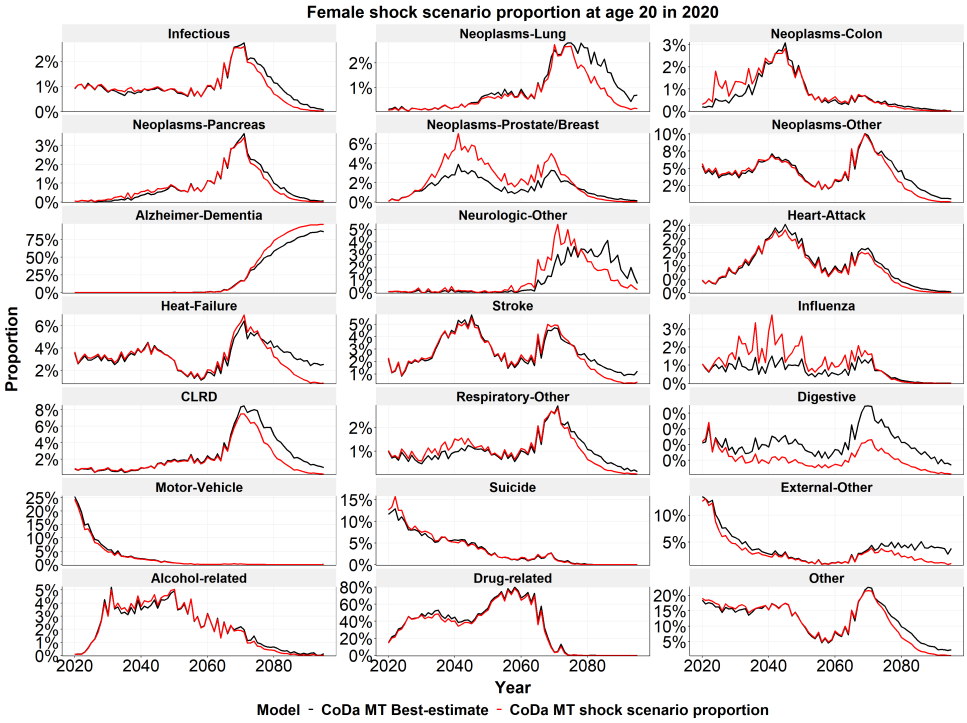


Figure 20: L'âge attendu du décès pour la trajectoire centrale et pour le scénario 0.5% pour une femme de 20 ans en 2020

Par conséquent, l'analyse devrait être approfondie en tenant compte des jugements d'experts, qui pourraient être appliqués à chaque cause de décès pour limiter leur évolution à une échelle plus rationnelle. Certains problèmes fondamentaux liés à la modélisation des causes de décès restent également à résoudre, comme la qualité des données suite au changement de norme de classification.

En conclusion, l'application de la modélisation des causes de décès est jugée prématurée sans ajustement supplémentaire et dans le cadre de l'évaluation interne du risque de mortalité. Cependant, elle permet de fournir des informations précieuses sur les tendances et sur l'évolution de la mortalité, de détecter les principaux facteurs du risque de mortalité et d'analyser des scénarios extrêmes.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon alternance et qui m'ont aidé lors de la rédaction de ce mémoire, notamment l'ensemble des membres de l'équipe Model Analysis de SCOR pour leur encouragement pendant mon alternance.

Je voudrais remercier en particulier mes tuteurs Keunoo Chang et Philippe Bertin pour leur tolérance, patience et conseil, ainsi que leur expertise tout au long de ce mémoire, sans eux ce mémoire ne pourrait jamais être réalisé.

Je tiens également à témoigner toute ma reconnaissance à Julien Tomas, actuaire R&D chez SCOR, pour son partage généreux de connaissances sur le modèle CoDa et sa revue de ce mémoire.

Enfin, je voudrais remercier mes parents pour leur soutien inconditionnel depuis toujours.

Contents

Introduction

Mortality rates has always been the focus of life insurance as well as demographic researcher. For pension funds and government, an accurate mortality rates prediction impact the decision making regarding various social issues such as retirement policy reform; social resources allocation. As for life insurance companies, mortality rates provide a solid guideline for both policy pricing and risk management, notably compliance with Solvency II regulation, which requires strict quantitative risk assessment aiming to deduce adequate Solvency Capital Requirement (SCR). According to Solvency regulation in force, two options are available for European(re-)insurance companies to evaluate their capital in Pillar I, either a standard formula along with the regulation or an internal model developed by the insurer, which proposes numerous advantages especially accuracy on the real risk encountered by the insurer.

Hence, mortality modelling is vital for an insurance company with regard to its profitability and risk resilience capacity although current modelling practice based on All-Cause information could be further refined with respect to cause-specific experiences among both young and old ages.

The common practice of the mortality modelling is to consider 3 dimensions of variability: *Gender; Age* and *Temporality*. In light of the current situation, a cause-of-death modelling approach could be insightful to capture the uncertainty within the mortality evolution in a more granular way. Cause-of-death approach reckons with the source of death as well as its development. This report aims to apply this modelling approach and compare it firstly with the common practice classic mortality forecast and subsequently evaluate each cause-of-death model's suitability in a mortality risk framework, and their own advantages and limits. This analysis is performed on the US general population, the result could differ according to each insurer's portfolio.

This thesis will start by reviewing regulation on mortality risk and cause-of-death definitions in the first chapters. Third chapter aims to apply a model based on independent cause assumption as a first tentative, each cause is modelled and forecasted independently, the aggregation is realized by summing the forecast of each cause. The fourth chapter will explore compositional data analysis techniques on cause-of-death modelling in order to further emphasize the dependence structure between causes and provide more coherent cause-specific forecast.

Chapter 1

Study context

This chapter introduces the background and motivation of this thesis regarding mortality risk modelling.

1.1 Reinsurance

Reinsurance could be directly interpreted as insurance of insurance, according to the type of insurance purchased by the cedant (primary insurer), the reinsurance company accepts part or total of the risks encountered by an insurance company in exchange for reinsurance premiums. Insurance companies could benefit from this mechanism of risk transfer to avoid large exposure to an unforeseen event which leads to massive claims, and in the meantime obtain more underwriting capacity without damaging their solvency. Reinsurance company, on the other hand, ends up with a more diversified portfolio which mitigates its own risk as well. If a reinsurance company estimates that the total risk ceded in its portfolio may surpass its capacity, the ceded risk could be further shared with other reinsurance companies with retrocessions. There are mainly two types of reinsurance: proportional and non-proportional.

Proportional reinsurance

The part of risk transferred to the reinsurance company is equal to the proportion of ceded premiums. The determination of the proportion could be done in two types:

Quote-share (QS): cedant and reinsurance company mutually agree on a fixed unique rate for each policy.

Surplus-share: cedant retains a limit of the proportion of losses (retention) and transfers the excess of proportions to the reinsurance company.

Non-proportional reinsurance

Non-proportional reinsurance is mainly executed in two forms: excess of loss and stop-loss.

Excess of loss indicates a lower and upper threshold of claims, cedant retains the part below the lower threshold and above the upper threshold. The rest is taken charge by the reinsurance company.

Stop-loss reckons with the annual result of cedant, reinsurance company takes over the part of loss above the pre-negotiated limit.

Non-proportional reinsurance is mainly used in event-based policies such as natural catastrophe claims.

1.2 Regulation landscape

1.2.1 Solvency II framework

The solvency of (re-)insurance companies is defined as the ability to meet their short, medium and long-term commitments to their clients. It depends on the size of these commitments, including the guarantees and protection offered to the (re-)insured, and the resources available to the (re-) insurance company to meet them (equity and assets held).

With the purpose of enhancing risk resilience and unifying solvency regulation across European (re-)insurance companies. Solvency II came into effect in January 2016 and could be characterized by 4 features:

- Market consistent: valuation methodologies should be based on the exchangeable value in the market.
- Risk-based: Capital requirement in consistence with the risks hold in the portfolio, higher risks undertaken by (re-)insurance companies require higher capital requirement.
- Proportionate: Appropriate regulatory conditions corresponding to the nature, scale and complexity of the risks inherent in the insurance and reinsurance business.
- Group supervision: Supervisors should strengthen the internal coordination and information exchange of the Supervisory Committee.

Based on the characteristics and objectives mentioned above, Solvency II executes mainly through a structure of 3 pillars:

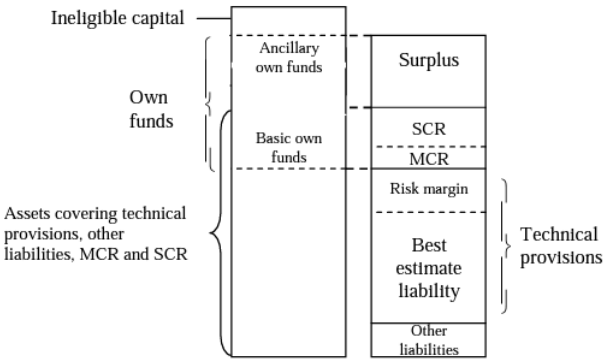


Figure 1.1: Pillar I
Source: Institute and Faculty of Actuaries 2016

Pillar I requires each (re-) insurer to appropriately quantify their technical provisions valuation with reliable specific methodologies, which result in Minimum Capital Requirement (MCR) and Solvency Capital Requirement (SCR). MCR and SCR are both capital requirements above the technical provisions. MCR is the requirement of a minimum capital level that (re-)insurance companies should hold and is regularly supervised by the regulatory authority. SCR corresponds to the Value-at-Risk at 99.5% of future companies' own funds, which limits the probability that (re-)insurance companies can't meet their obligations towards policyholders under 0.5% and mitigate the risk of financial ruin in one year.

Pillar II imposes qualitative conditions on each (re-)insurer's risk management unit, which is realized by the Own Risk and Solvency Assessment (ORSA), to identify types of

risk to which (re-insurance) companies are exposed and oversight ongoing risk management process and controls

Pillar III consists of a risk reporting system according to which risk assessment shall be reported to regulators and the public.

Figure 1.2: Solvency II review in 2020
Source:European commission

It should be noted that Solvency II, despite that it has come into effect, still has been through several revisions and updates in order to ensure its stability regarding its objectives. EIOPA (the European Insurance and Occupational Pensions Authority) accordingly produced corresponding technical support and advice. The most recent review took place in 2020 and EIOPA is of opinion that no major changes in Solvency II framework are needed, Figure 1.2 shows the three areas of improvement provided by EIOPA.

Regular reviews of Solvency II also provide aspects to (re-)insurance companies to adapt their own risk assessment methodologies with regard to the slowing economic growth, and uncertainties regarding market conditions which are intensified by the Covid-19 pandemic.

1.2.2 Standard formula

As in pillar I of the Solvency II directive, risk assessment and quantification may be accomplished by a standard formula, which prescribes the stress tests or methodology of aggregation techniques. Figure 1.3 indicates the structure of SCR under the standard formula. The basic SCR is calculated in each individual module.

The SCR is composed of three components: Basic SCR; Operational risk and Adjustment. Basic SCR is further divided into different modules: market (interest rate; credit spread; currency etc); counterparty default; intangible assets and insurance risks (including Health, Life and Non-life risks). Among these, life underwriting risk includes biometric risks, which may cause (re-)insurance companies large claims due to human life conditions (death, disability, birth etc), it includes mortality risk; longevity risk and disability/morbidity risk.

Mortality risk refers to the risk that the actual payments arising from policyholder deaths, during the term of the cover, exceed the expected payments as a result of mortality experience being higher than expected. This risk is defined in the standard formula as a permanent 15% increase in mortality rates. Apart from biometric risks, lapse, expenses, revision and catastrophe risk are also included in the calculation of life module basic SCR.

The basic SCR of each individual risk is calculated as the difference between the central scenario balance sheet and the stressed balance sheet, which is further combined within each module. Having obtained the basic SCR for each module, they are aggregated

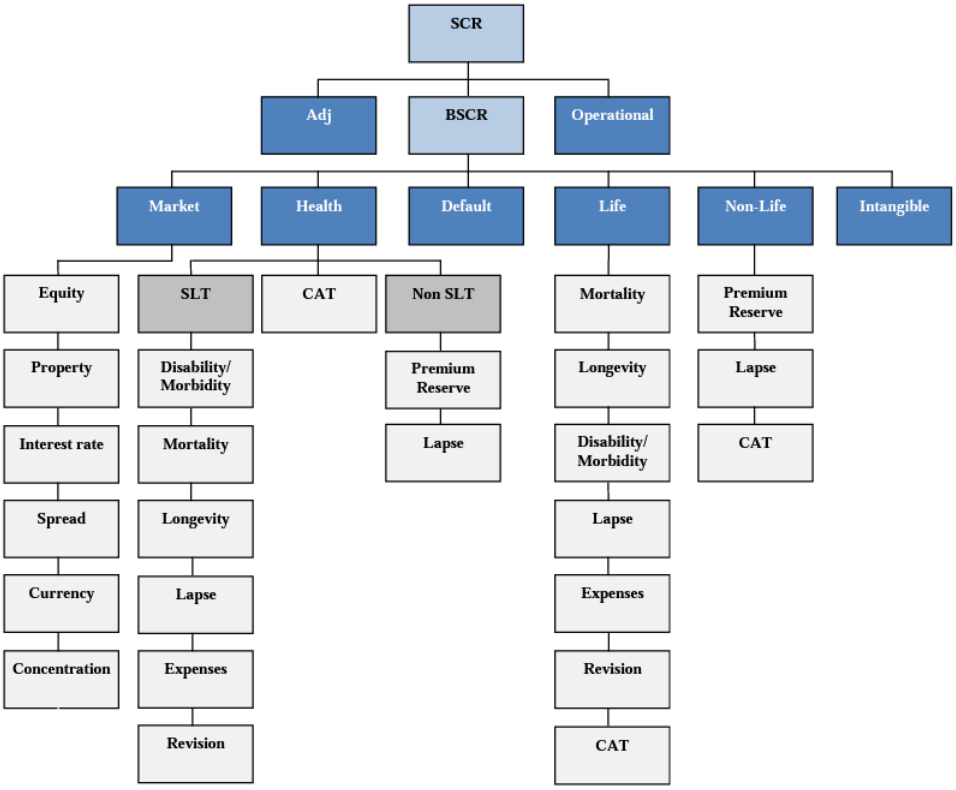


Figure 1.3: Solvency capital requirement branch
Source: Institute and Faculty of Actuaries2016

using a specific correlation matrix at different levels

$$SCR_{Life} = \sqrt{\sum_{i,j} \rho_{i,j} \times SCR_i \times SCR_j}$$

$\rho_{i,j}$ represents the correlation between two individual risks

$$BSCR = \sqrt{\sum_{I,J} \rho_{I,J} \times SCR_I \times SCR_J}$$

$\rho_{I,J}$ represents the correlation between two modules

In the end, basic SCR is completed by an allowance to operational risk and an adjustment regarding the loss-absorbing capacity. Adj corresponds to the adjustment for loss-absorbing capacity of technical provisions and deferred taxes, which are not taken into account in the BSCR. Operational risk refers to the loss risk due to the internal procedure, inadequate systems or external events.

Standard formula possesses advantages, it is less costly and easy to implement. Although it doesn't provide a customized view of risks encountered by each (re-)insurance company. Furthermore as mentioned before, regular reviews of Solvency II motivates (re-)insurance companies to adapt their risk assessment as well. Therefore an alternative approach accentuating (re-)insurance companies' own risk profile may be insightful.

1.2.3 Internal model

Each (re-)insurance company possesses a different risk profile, (re-) insurance companies have the right to develop its own internal model under the approval of its supervisory authority. The advantage of internal model is that it sheds light on each (re-)insurance

company's individual characteristics. The use of internal model is required to meet several standards such as

- Use test: proof of internal model impacts on company decision making and governance process
- Statistical soundness: minimum statistical standards must be met, as well as justification of methodology: expert judgement and aggregation method etc. Use test is vital for a (re-) insurance company to receive approval on the use of its internal model.
- Calibration standard: analysis of whether the SCR correspond to 1-200 scenario
- Profit & loss distribution: justify real profit of loss by the categorisation of risk chosen in the internal model
- Validation: Regular review and validation

The use of an internal model may also be subject to issues such as data quality and assumptions such as the determination of the calibration period, the choice of calibration period may limit the understanding of extreme events due to the length of data or the volatility cluster as in the financial market.

The advantage of the use of the internal model is to assess appropriately each (re-)insurance company's risk profile, which is also an aspect urged by regulatory authorities on capturing correctly the exposure, allowing for more flexibility on its own funds and gaining more precision on the risk modelling. According to PwC [2019], between 2017 and 2018, internal models helped (re-)insurers to reduce on average 8.7% of SCR compared to the Standard formula.

Internal models commonly use simulation methods such as Monte-Carlo to project (re-)insurance companies' future profit and loss distribution, which consists of simulating future possible scenarios of each individual risk and then aggregating them according to their aggregation techniques.

1.3 Mortality evolution in US

Mortality rates have been through a rapid improvement since World War II, as well as life expectancy which measures the average life that an individual at an attained age expects to live. Life expectancy at birth in the US raised from 68.2 to 78.7 all races and genders included according to Elizabeth Arias and Ahmad [2022], this improvement mainly benefits from medical advances and a relatively stable environment. Nonetheless, it has come to public attention that US life expectancy at birth declined for the first time in two decades in 2015 and subsequently in 2016 and 2017. Despite a slight recovery until 2019, the public health crisis COVID-19 brings life expectancy to its lowest level since 1996, which explained 50% of the decline between 2020 and 2021, along with other major causes such as unintentional injuries and heart diseases. The impact would've been greater without the compensation from the reduction of Influenza; Alzheimer and perinatal conditions.

Figure 1.4 illustrates the mortality rate evolution of a male aged 30 in the US from 1933 to 2019. Similarly to life expectancy, mortality rates demonstrate more uncertainty among the young and middle ages, National Academies of Sciences et al. [2021] states that the distortion among the working-age group is mainly due to drug overdose; suicide and cardio-metabolic diseases. It also has been established that the gap in life expectancy

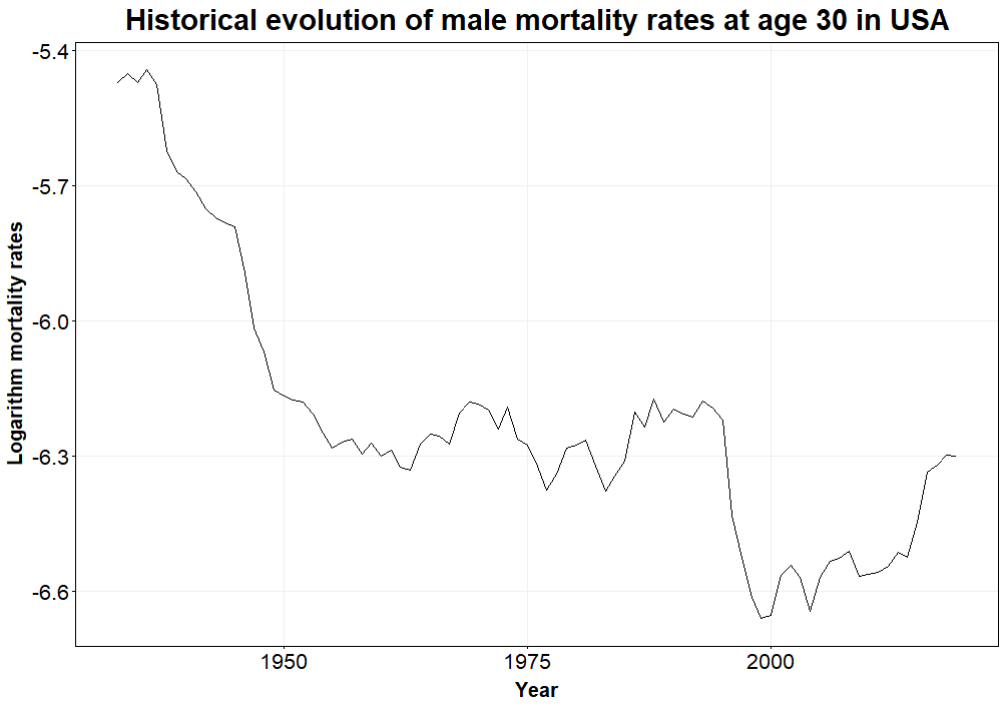


Figure 1.4: US 30-year-old male logarithm mortality rates between 1933 and 2019. *Data source: Human Mortality Database (HMD)*

induced by socioeconomic factors has been expanding according to Brown and McDaid [2003] and National Academies of Sciences and Medicine [2015], mainly due to education level and income.

The above issues reveal the major factor behind the apparent phenomenon could be better understood through cause-of-death analysis, an aggregate level mortality improvement potentially concealed the underlying cause-specific evolution.

Chapter 2

Mortality modelling approach

This chapter aims to introduce basic concepts related to mortality quantification and the modelling approach commonly practised in both the research and the insurance industry.

2.1 Mortality notations

Before tackling the modelling approach to mortality, several mortality topic-related notations ought to be clarified. By denoting individual lifetime as T and residual lifetime as T_x, survival and mortality probability could be defined:

$$
\begin{aligned}
{}_tp_x &= P(T_x > t) = P(T > x + t | T > x) \\
{}_tq_x &= P(T_x \leq t) = P(T \leq x + t | T > x)
\end{aligned}
$$

Another notion of mortality force μ_{x+t}, which measures the instantaneous conditional death probability, is deduced directly from the definition above and could be related to mortality rate, as per definition:

$$
\mu_{x+t} = \lim_{h \to 0} h^{-1} P(t < T_x \leq t + h | T_x > t) = \frac{1}{{}_tp_x} * \frac{\partial_t q_x}{\partial t}
$$

A central mortality rate between x and $x + 1$ could be calculated as below:

$$
m_x = \frac{d_x}{n_x}
$$

where d_x represents the number of deaths within a 12-month period and n_x quantity of exposures defined as the average number of populations at age x between the beginning and end of the 12 months, this mortality rate could be easily calculated according to the chosen mortality database such Human Mortality Database (HMD).

μ_x is therefore linked with m_x, $\mu_x = \lim_{h \to 0} m_x$ and the mortality probability could be consequently written and approximated as:

$$
{}_tq_x = 1 - exp(- \int_0^t \mu_{x+s} \, ds) = \frac{\mu_x}{1 + 0.5\mu_x}
$$

Life expectancy

Complete life expectancy at age x could be expressed as $\dot{e}_x = E(T_x) = \int_0^\infty {}_tp_x dt$. Curtate life expectancy e_x is the expectation of discrete random variable K_x which is the integer

part of T_x, it represents the complete years of life count.

$$e_x = E(K_x) = \sum_{k=1}^{\infty} {}_kp_x$$
$$\dot{e}_x \approx e_x + 0.5$$

Life expectancy may be calculated from two types of life tables:

Period life table or cohort table. The period table calculates the mortality rates from a single year and assumes that for the rest of life, the mortality rates will remain the same. The period of life expectancy could be expressed as :

$${}_kp_x^{Period} = p_x(t) \times p_{x+1}(t) \times p_{x+2}(t) \times ... \times p_{x+k}(t)$$
$$\dot{e}_x^{Period} = E(K_x) = \sum_{k=1}^{\infty} {}_kp_x^{Period} + 0.5$$

The cohort life table provides the death rates of a virtual cohort, it takes into account the mortality improvements in the future, therefore the survival probabilities used in the cohort life expectancy for an individual aged 20 in 2020 will be the survival probabilities of age 21 in 2021, age 22 in 2022 ..... age 60 in 2060 etc, thus the resulting life expectancy is:

$${}_kp_x^{Cohort} = p_x(t) \times p_{x+1}(t+1) \times p_{x+2}(t+2) \times ... \times p_{x+k}(t+k)$$
$$\dot{e}_x^{Cohort} = E(K_x) = \sum_{k=1}^{\infty} {}_kp_x^{Cohort} + 0.5$$

2.2 Stochastic mortality model

2.2.1 Lee-Carter

Lee-Carter model is a widely used mortality stochastic model in both industry and research fields, first introduced by Lee and Carter [1992]. It models two dimensions of mortality: age and time through three terms: α_x indicates a static age structure, κ_t captures the time dynamics notably the trend of mortality, β_x an interaction term between age and time dynamics measures the sensitivity of each age towards the general trend, each age undergoes different mortality evolution in terms of sign and magnitude.

$$\ln \mu_{x,t} = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}$$

With constraint $\sum_t \kappa_t = 0$ and $\sum_x \beta_x = 1$,the parameter estimation initially presented in Lee and Carter [1992] is accomplished by singular value decomposition (SVD). Given a matrix of logarithms of mortality rates with dimension N ages x T years, the matrix is initially extracted by its age-specific rates α_x and the rest is decomposed by SVD:

$$Z_{x,t} = (ln(\mu_{x,t}) - \alpha_x)$$
$$Z_{x,t} = \sum_{i}^{T} \rho_i U_{x,i} V_{i,t}$$

with T as the total number of ranks and $U_{x,i}$and $V_{i,t}$ as the left and right singular vectors, it follows by a low-rank approximation of rank h Eckart and Young [1936], where h is the h first ranks chosen in T.

$$\widehat{Z_{x,t}} = \sum_{i}^{h} \rho_i U_{x,i} V_{i,t} = \sum_{i}^{h} \beta_x^i \kappa_t^i$$

Despite the fact that Lee and Carter [1992] proposed an approximation by rank 1, the choice of rank could also be determined by the total explained variance of h-rank approximation: $\sum_i^h \rho_i^2$.

Forecast

Forecast of Lee-Carter model is accomplished by extrapolating κ_t, after α_x, κ_t and β_x are estimated as described above, κ_t is then further modelled using time series techniques, a random walk with drift model provides a satisfactory result.

$$\kappa_t = \kappa_{t-1} + \delta + \epsilon_t$$
$$\widehat{\kappa_{t+1}} = \kappa_t + \delta$$

One of the limits presented by Lee-Carter model is its lack of consideration of the medical improvement or environmental change, this model emphasizes the historical observation from which the pattern and trend observed are assumed to continue in the future by the extrapolation of time-varying factor κ_t,

2.2.2 Poisson log-bilinear

An extension to Lee-Carter model is Poisson log-bilinear model of Brouhns et al. [2002], compared to Lee-Carter model, this model successfully addressed the unseemingly realistic hypothesis of homoscedasticity of $\epsilon_{x,t}$, due to higher volatility induced by the limited exposure numbers of extreme age.

The main difference is to model the number of deaths $D_{x,t}$ as a random variable of Poisson distribution:

$$\mu_{x,t} = exp(\alpha_x + \beta_x \kappa_t)$$
$$D_{x,t} \sim Poisson(E_{x,t} * \mu_{x,t})$$

$E_{x,t}$ represents as before the exposure and $m_{x,t}$ the mortality rate modelled by Lee-Carter model. Instead of a SVD decomposition method, Poisson log-bilinear model estimation is accomplished by Maximum Log-Likelihood and holds the same parameters constraint regarding κ_t and β_x.

The log-likelihood function of the Poisson log-bilinear model could be written as:
$L(\alpha, \beta, \kappa) = \sum_{x,t}[D_{x,t}(\alpha_x + \beta_x \kappa_t) - E_{x,t} exp(\alpha_x + \beta_x \kappa_t)] + constant.$

Newton iterative method is implemented to update the estimation of the parameters θ:
$\widehat{\theta^{(v+1)}} = \widehat{\theta^{(v)}} - \frac{\partial L^{(v)}/\partial \theta}{\partial^2 L^{(v)}/\partial \theta^2}$

One of the three sets of parameters θ: $\alpha_x; \beta_x; \kappa_t$ is updated at each iterative v. The interpretation of parameters remains the same with Lee-Carter model, κ_t is assumed to follow an ARIMA (0,1,0) and linearly extrapolated to forecast future time dynamics.

$$\kappa_t = \kappa_{t-1} + \delta + \epsilon_t$$

A first application on HMD US male mortality data

Figure 2.4 illustrates a first tentative to model of male logarithm mortality force in US from 2000 to 2019, α_x shows the general pattern of mortality force over different ages, after peak mortality rates due to the infant mortality, mortality force demonstrates a log-linear shape over ages, especially after age 20. κ_t proves the overall mortality improvement among the general population, which is in line with the increase in life expectancy observed after World War II. Despite the fact that each age illustrates distinct sensitivity towards the overall mortality improvement, the majority of age groups act in accordance with the common trend (positive β_x). Nonetheless, it also appears that males aged between 25 - 30 in US detached from the improvement (negative β_x), furthermore at the elder ages, this improvement seems to decline or stagnate, explication among extreme ages (age > 95) may be difficult due to the volatility and wake exposure numbers, but these phenomenons are in the interest of this thesis to understand from a cause-of-death view.

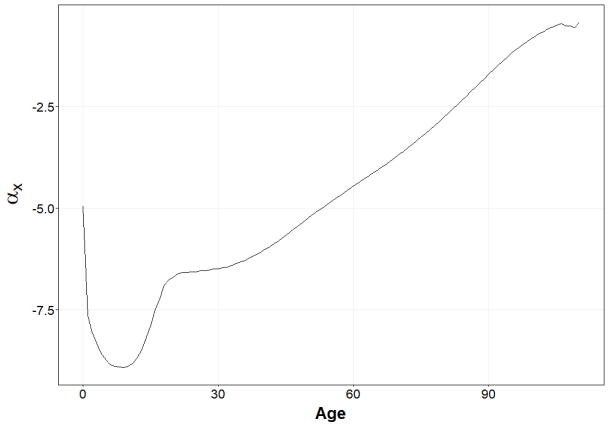


Figure 2.1: α_x

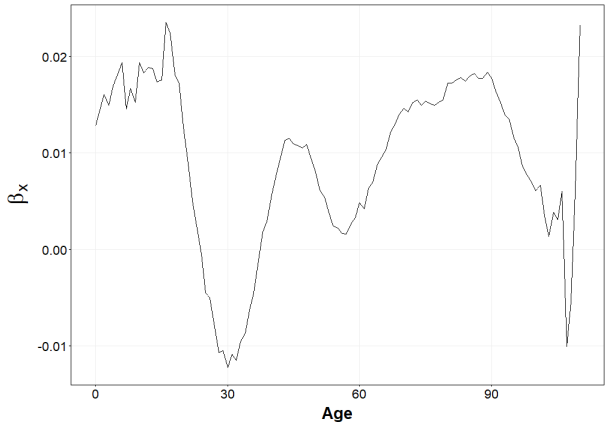


Figure 2.2: β_x

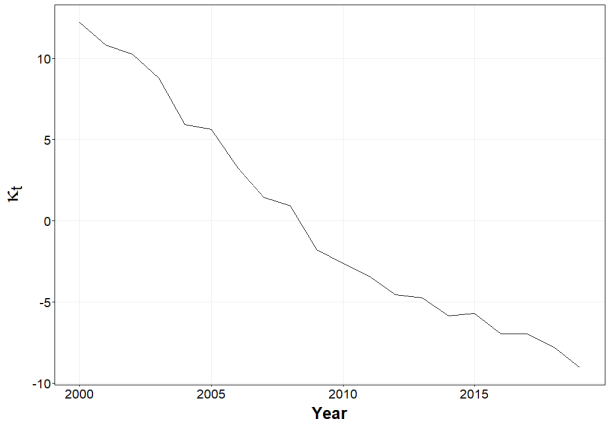


Figure 2.3: κ_t

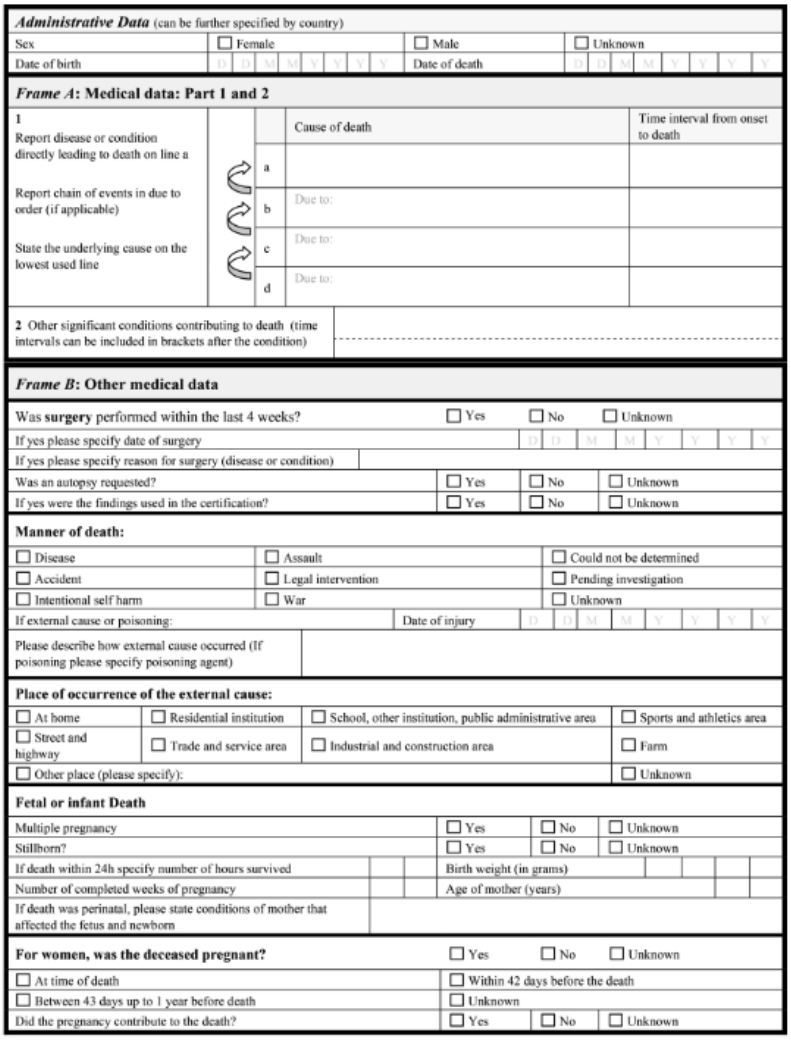
Figure 2.4: Poisson log-bilinear model parameters on US male logarithm mortality force, data source: Human Mortality Database (HMD) US mortality data 2000-2019

2.3 Cause-of-death: definition; classification and data

2.3.1 Cause-of-death definition

The underlying cause of death is defined as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or

violence which produced the fatal injury" according to the World Health Organization (WHO), which is designated in the death certificate (template shown in 2.5).



Administrative Data (can be further specified by country)

Sex	☐ Female	☐ Male	☐ Unknown
Date of birth	D D M M Y Y Y Y	Date of death	D D M M Y Y Y Y

***Frame A*: Medical data: Part 1 and 2**

1 Report disease or condition directly leading to death on line a		Cause of death	Time interval from onset to death
Report chain of events in due to order (if applicable)	a		
	b	Due to:	
State the underlying cause on the lowest used line	c	Due to:	
	d	Due to:	

2 Other significant conditions contributing to death (time intervals can be included in brackets after the condition)

***Frame B*: Other medical data**

Was **surgery** performed within the last 4 weeks?	☐ Yes	☐ No	☐ Unknown
If yes please specify date of surgery	D D M M Y Y Y Y		
If yes please specify reason for surgery (disease or condition)			
Was an autopsy requested?	☐ Yes	☐ No	☐ Unknown
If yes were the findings used in the certification?	☐ Yes	☐ No	☐ Unknown

Manner of death:

☐ Disease	☐ Assault	☐ Could not be determined
☐ Accident	☐ Legal intervention	☐ Pending investigation
☐ Intentional self harm	☐ War	☐ Unknown
If external cause or poisoning:	Date of injury	D D M M Y Y Y Y
Please describe how external cause occurred (If poisoning please specify poisoning agent)		

Place of occurrence of the external cause:

☐ At home	☐ Residential institution	☐ School, other institution, public administrative area	☐ Sports and athletics area
☐ Street and highway	☐ Trade and service area	☐ Industrial and construction area	☐ Farm
☐ Other place (please specify):			☐ Unknown

Fetal or infant Death

Multiple pregnancy	☐ Yes	☐ No	☐ Unknown
Stillborn?	☐ Yes	☐ No	☐ Unknown
If death within 24h specify number of hours survived		Birth weight (in grams)	
Number of completed weeks of pregnancy		Age of mother (years)	
If death was perinatal, please state conditions of mother that affected the fetus and newborn			

For women, was the deceased pregnant?	☐ Yes	☐ No	☐ Unknown
☐ At time of death	☐ Within 42 days before the death		
☐ Between 43 days up to 1 year before death	☐ Unknown		
Did the pregnancy contribute to the death?	☐ Yes	☐ No	☐ Unknown

Figure 2.5: International form of medical certificate of cause of death provided by WHO

Nevertheless, the definition of each cause of death may vary among different countries and regions, which brings the challenge of the mortality reporting system and worldwide study on the cause of death study, from which was initiated International classification of diseases (ICD), appeared in the 19th century and latest updated after the 11th revision in 2019, it came into effect on January 2022. The ICD serves as the broad standard of classification of death, each disease is classified with a unique code in order to facilitate further use, most cause-of-death databases stick to the same standard (code).

However, it is complicated to work with more than a hundred causes of death in an actuarial context, necessary mapping into a group of causes is then proceeded according to the pathology; medical expert opinion and its impact on human health. 21 causes of death are considered in total.

The 21 causes listed in Table 2.1 cover major causes of death and its mapping is deemed appropriate, especially regarding the target country of this report: US.

2.3.2 Cause-of-death public databases

Human Mortality Database (HMD) is the current most widely used database in mortality modelling. Several public US general population cause-of-death databases are available

Number	Cause name
1	Infectious
2	Neoplasms-Lung
3	Neoplasms-Colon
4	Neoplasms-Pancreas
5	Neoplasms-Prostate/Breast
6	Neoplasms-Other
7	Alzheimer-Dementia
8	Neurologic-Other
9	Heart Attack
10	Heart-Failure
11	Stroke
12	Influenza
13	Chronic Lower Respiratory Disease (CLRD)
14	Respiratory-Other
15	Digestive
16	Motor-Vehicle
17	Suicide
18	External-Other
19	Alcohol-related
20	Drug-related
21	Other

Table 2.1: Cause-of-death mapping classification list

with open access such as Human cause-of-death database (HCD) and Centers of Disease Control and Prevention (CDC) data. Human cause-of-death database (HCD) provides a number of deaths by cause in each five-year age range are the most similar to HMD, the total number of deaths is consistent with HMD. An advantage of HCD is that it takes into account the impact of classification version changes and thrives to provide more consistent cause-of-death numbers according to a constant classification, it facilitates the investigation of the study of cause-specific trends.

death	country	year	sex	list	agf	cause	total	d0	d95p
1	US	1979	1	interm	4	0	1044958.00	25996.86	10027.86
2	US	1979	1	interm	4	1	171.21	24.02	2.70
3	US	1979	1	interm	4	2	371.73	63.12	5.25
4	US	1979	1	interm	4	3	1538.68	0.95	4.41
5	US	1979	1	interm	4	4	5601.86	182.43	78.99
6	US	1979	1	interm	4	5	622.51	67.44	2.91
7	US	1979	1	interm	4	6	0.00	0.00	0.00
105	US	1980	1	interm	4	0	1075078.00	25808.69	11415.13
106	US	1980	1	interm	4	1	189.49	22.46	3.15
107	US	1980	1	interm	4	2	383.15	56.04	9.35
108	US	1980	1	interm	4	3	1488.61	2.02	9.39

Table 2.2: Example of HCD data

	Year	Sex	Education	Age	COD113.lvl.1	COD113.lvl.2	COD113.lvl.3	COD113.lvl.4	COD113.lvl.5	Deaths
1	2008	M	College	98	9	9	9	9	9	0
2	2002	M	Unknown	98	9	9	9	9	9	0
3	2007	M	Graduate	98	9	9	9	9	9	0
4	2002	M	College	98	9	9	9	9	9	0
5	2005	M	Bachelor	98	9	9	9	9	9	0
6	2006	M	College	29	44	44	44	44	44	2
7	2001	F	Graduate	68	53	54	58	61	62	9
8	2002	M	High School	29	76	77	77	77	77	1
9	2009	M	College	29	76	77	77	77	77	2
10	2012	M	Unknown	29	76	77	77	77	77	0

Table 2.3: Example of CDC data

Table 2.3 provides an example of Centers of Disease Control and Prevention (CDC) data, which does not limit to providing cause-of-death data, and relevant periodic mortality reports involving cause-of-death, the census of the number of death of CDC is directly retrieved or reported from medical death certificates from all the states of US, each death certificate provides information about the single underlying cause-of-death and demographic data. Underlying cause-of-death is expressed as 4-digits ICD codes (110 causes of death, 113 selected causes of death, 130 selected causes of death for infants, and categories for injury intent and mechanism, or drug/alcohol-induced causes of death)

US mortality data produced by CDC are treated with final adjustment, another advantage of CDC data is the number of variables presented in the data, such as Educational level. It allows to perform flexible selection among the population to take into account the different mortality rates evolution in the sub-population. Table 2.4 summarises the respective features and advantages of each database.

	HCD	CDC
Data source	National Center for Health Statistics (NCHS)	- National Center for Health Statistics (NCHS) - Centers of Disease Control and Prevention (CDC)
Age format	Age groups : 0,1-4, 5,...95+	Single-year age: 0 - 110+
Cause-of-death classification	- Constant classification based on 3-digit ICD-10 codes - Provides mapping list: short and intermediary list in death numbers dataset, 103 causes in intermediary list.	- 4-digit ICD-10 - 113 selected causes of death - 130 selected causes of death for infants
Data availability	1979 - 2018	(Most recent) 1999-2020
Advantages	- Similarity with HMD - Stable classification	- More genuine data source - More variable choices which allow to study different sub-populations

Table 2.4: Summary of available US general population mortality database

2.3.3 Adjustments

This thesis used CDC data from 2000 to 2019 in order to ensure stability regarding some variables and the version of ICD applied in the database. Furthermore, the main study age range in this thesis is 20-95, in line with the main age range used in the current actuarial context. A preliminary minor smoothing by adding 0.5 to the death numbers is also performed, in order to eliminate zero deaths of some causes in young or elder ages.

2.4 Literature review of Cause-of-Death modelling approach

Many researchers have studied cause-of-death modelling and it captures more attention in recent years given the mortality evolution issues shown above, some of them thrived to maintain the use of the classic stochastic mortality model under a multivariate form. Alai et al. [2018] suggested a multinomial logistic model capture the intrinsic competing risk nature among causes, Boumezoued et al. [2019] proposed a multivariate Lee-Carter framework to model independently the central trajectory (Best-estimate) of each cause of death by Lee-Carter model, and use a correlation matrix based on the residuals of κ_t to capture the dependence structure within causes.

Another approach inspired by Aitchison [1982] on Compositional data analysis (CoDa), which involves modelling non-negative components whose sum is constant, Oeppen et al. [2008] first applied CoDa on the multiple-decrement mortality model on the Japanese population, it models the life table death distribution to obtain more coherent cause-specific mortality. Afterwards Bergeron-Boucher et al. [2017] follows the same techniques on multi-population mortality forecast, it has been proved the dependence structure among different causes is directly incorporated in the central trajectory and conciliates the evolution of each cause with the aggregate level mortality rates. Kjærgaard et al. [2019] proposed two variant models on the base of Oeppen et al. [2008] with aim to consider specific characteristics within each cause. Instead of modelling life table deaths distribution, Piveteau and Tomas [2018] suggested imposing an aggregated mortality forecast constraint and model proportions of each cause. The CoDa application on mortality models especially in a cause-of-death context is still developing but has already displayed satisfactory results emphasising the competing risk concept and providing a coherent forecast of each cause of death.

Along with the trend of machine learning and the enhancement of computational power, Ludkovski et al. [2018] and Huynh and Ludkovski [2021] proposed a non-parametric way to model the dependence structure through the Gaussian process, based on the covariance function. This approach is still novel and not mature, and how to relate it to the context of mortality risk calibration still needs to be deepened.

This thesis first investigated the modelling approach proposed in Boumezoued et al. [2019] because of its convenient implementation and convenience to output results of mortality risk. After evaluating its performance and limits, Compositional data analysis (CoDa) models suggested in Oeppen et al. [2008]; Piveteau and Tomas [2018] is tested in order to address the limits encountered in the first model.

Chapter 3

Independent cause-specific model

This chapter intends to implement the multivariate Lee-Carter model as suggested in Boumezoued et al. [2019] with independent cause assumption and construct aggregate mortality rates prediction interval based on a correlation matrix calibrated between causes.

3.1 Theory and assumption

Firstly, aggregate level mortality force $\mu_{x,t}$ could be decomposed as the sum of cause-specific mortality force, indeed, given the number of deaths in each cause, the cause-specific mortality force which is assimilated by the mortality rates could be expressed as below.

$$\mu_{x,t,i} = \frac{D_{x,t,i}}{E_{x,t}}$$

$$\sum_i \mu_{x,t,i} = \mu_{x,t}$$

Each individual is ultimately dead of a single cause, not multiple causes at the same time, this is referred to as the competing risk framework, the lifetime of an individual could then be expressed as the minimum survival duration of each cause. For two causes A and B, the respective lifetime associated to each cause could be written as:

- T_A: lifetime of cause A
- T_B: lifetime of cause B
- T: lifetime
- $T = min(T_A, T_B)$

The lifetime of an individual is determined by the arrival of the first fatal cause. Therefore the survival function at the age a

$$S(a) = P(T_A > a, T_B > a) = exp(-\int_0^a \mu_y \, dy)$$

The net cause-specific probabilities is defined as $P(T_A < a + \delta | T_A > a)$ but not easy to be estimated in practice (**?** proposed a copula method to estimate net cause-specific intensities). Only the duration of the fatal cause can be observed while the duration of other causes are truncated, which means only the crude cause-specific probability $P(T_A < a + \delta | T = T_a, T_A > a)$ can be observed. Therefore it is convenient to assume causes are independent so that the cause-specific net rate could be determined by the crude death rate $\mu_{x,t,i}$, which leads to $S(a) = P(T_A > a) \times P(T_B > a) = exp(-\int_0^a \mu_A(y) \, dy) \times exp(-\int_0^a \mu_B(y) \, dy)$.

3.2 Historical observation

This section is on purpose to detect the historical trend of cause-of-death mortality, the historical trend in the calibration period plays a forceful role in the future forecast, particularly in the modelling of $\kappa_{t,i}$.

Figure 3.1, Figure 3.2 display the top 5 cause contributions and Figure 3.3, Figure 3.4 show the top 5 cause evolution within each age in the last year available in the data, it is apparent that different ages experience different major causes.

- As for the age range 20-50, *Drug* was revealed as a potential source of mortality improvement distortion due to its sharp increase. Not only it increased rapidly from 2000 to 2019, but also it appears as one of 5 major causes in some ages such as 50. *External* causes occupy an important role as well and causes related to cardiovascular diseases gain more influence starting from 40 years old.

- For age above 60, Cardiovascular diseases including *Hear-Attack* and *Heart-Failure* are the most important sources of death and the exposure probability to these two risks increases along with the age. More importantly, improvement among Cardiovascular causes seems to stagnate after the year 2010. On the other hand, *Dementia* impacts elder ages, especially above 85. Although, the sharp increase tends to stop after 2010.

- Age between 40 and 80 encounter more diversified causes since the sum of top 5 cause proportions are below other ages.

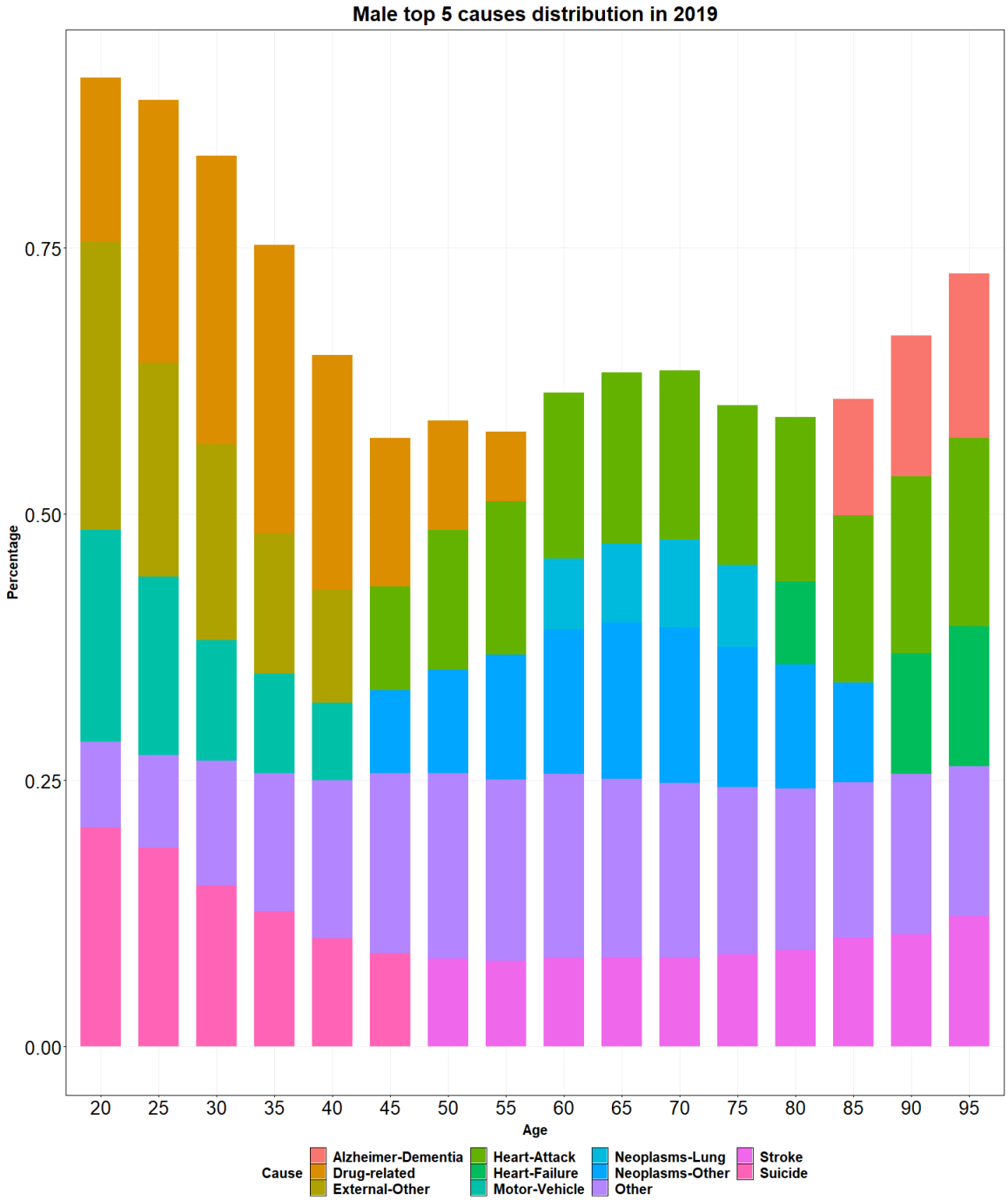


Figure 3.1: Top 5 male cause contributions by age

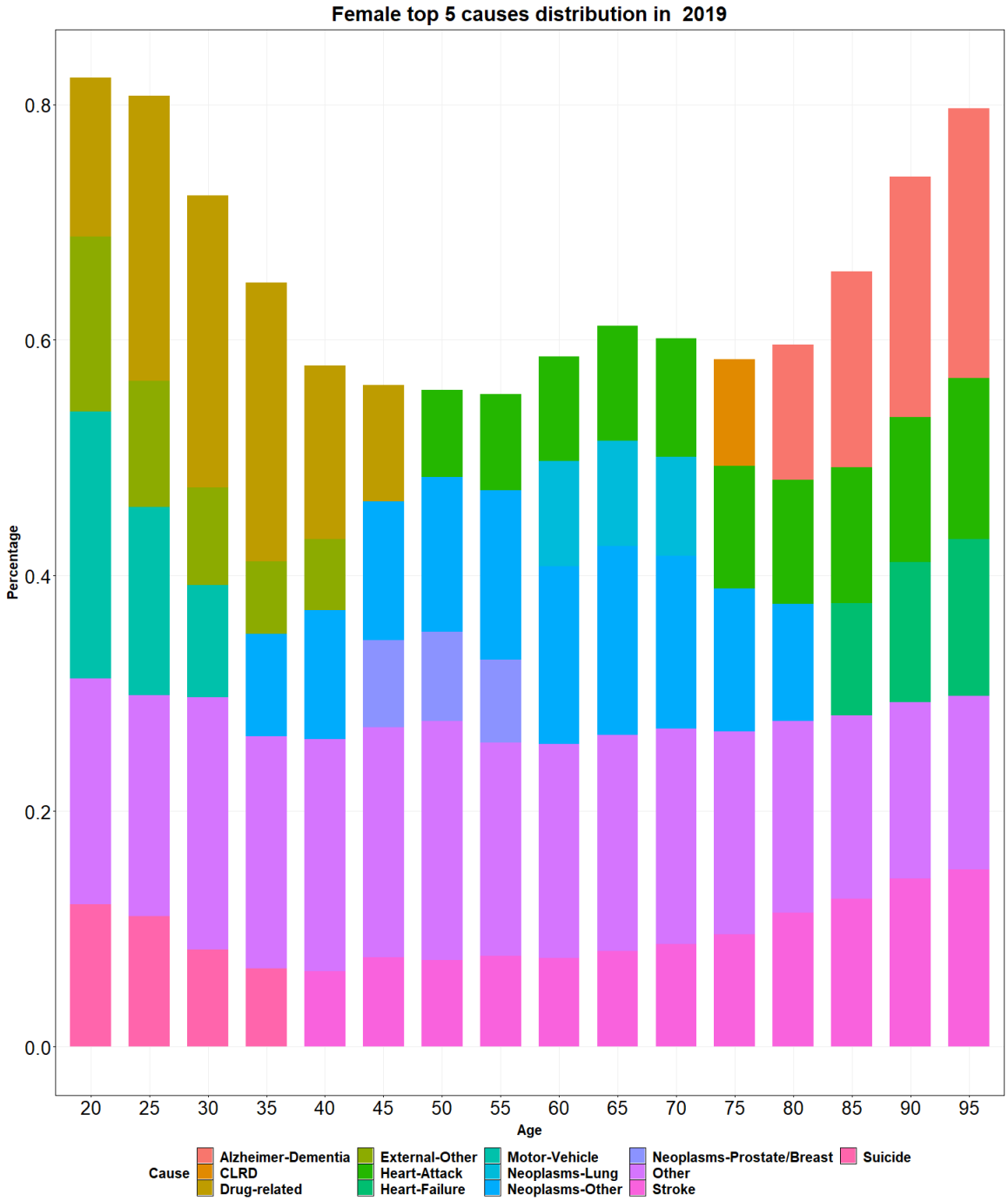


Figure 3.2: Top 5 female cause contributions by age

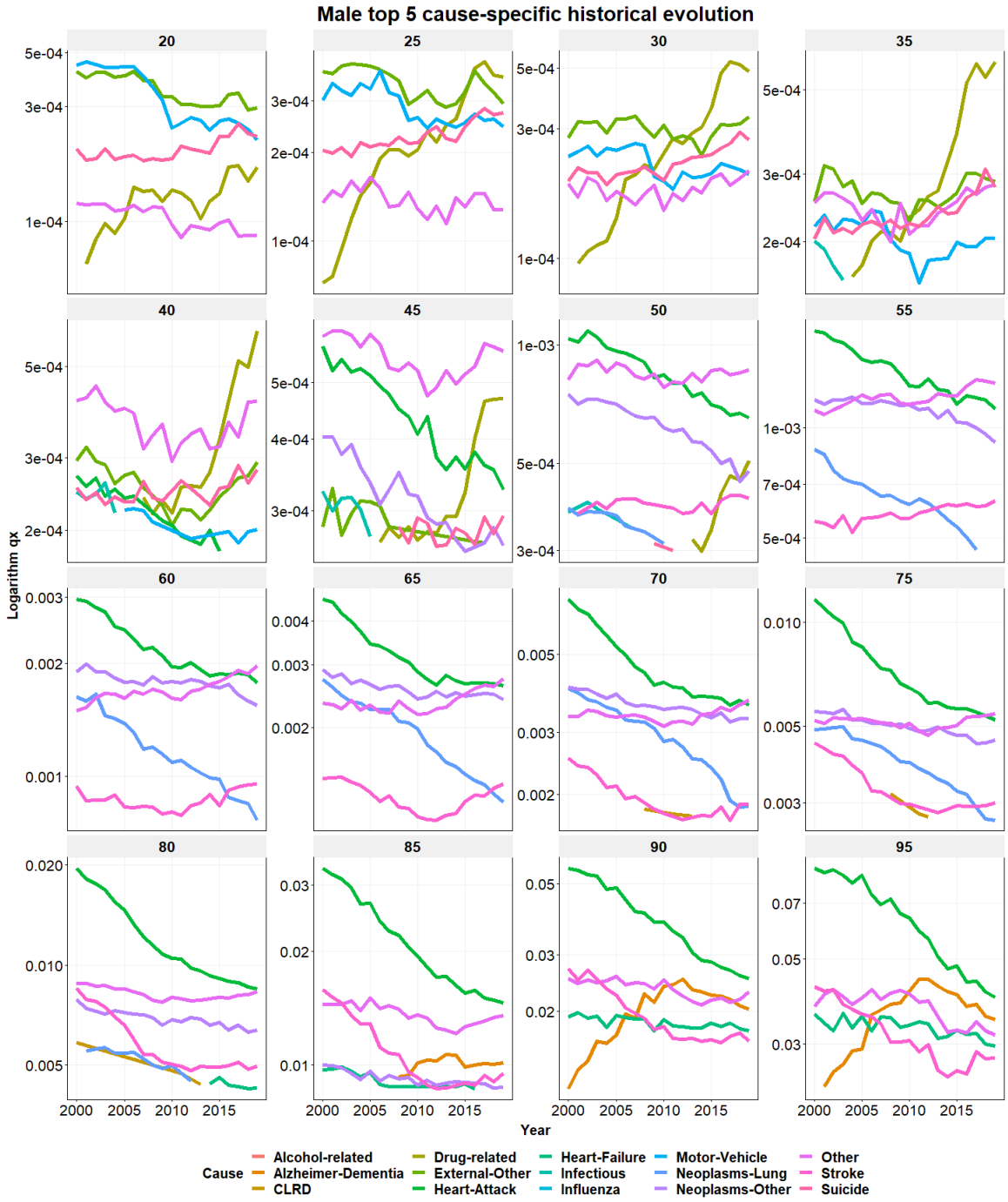


Figure 3.3: Top 5 male cause-specific historical evolution

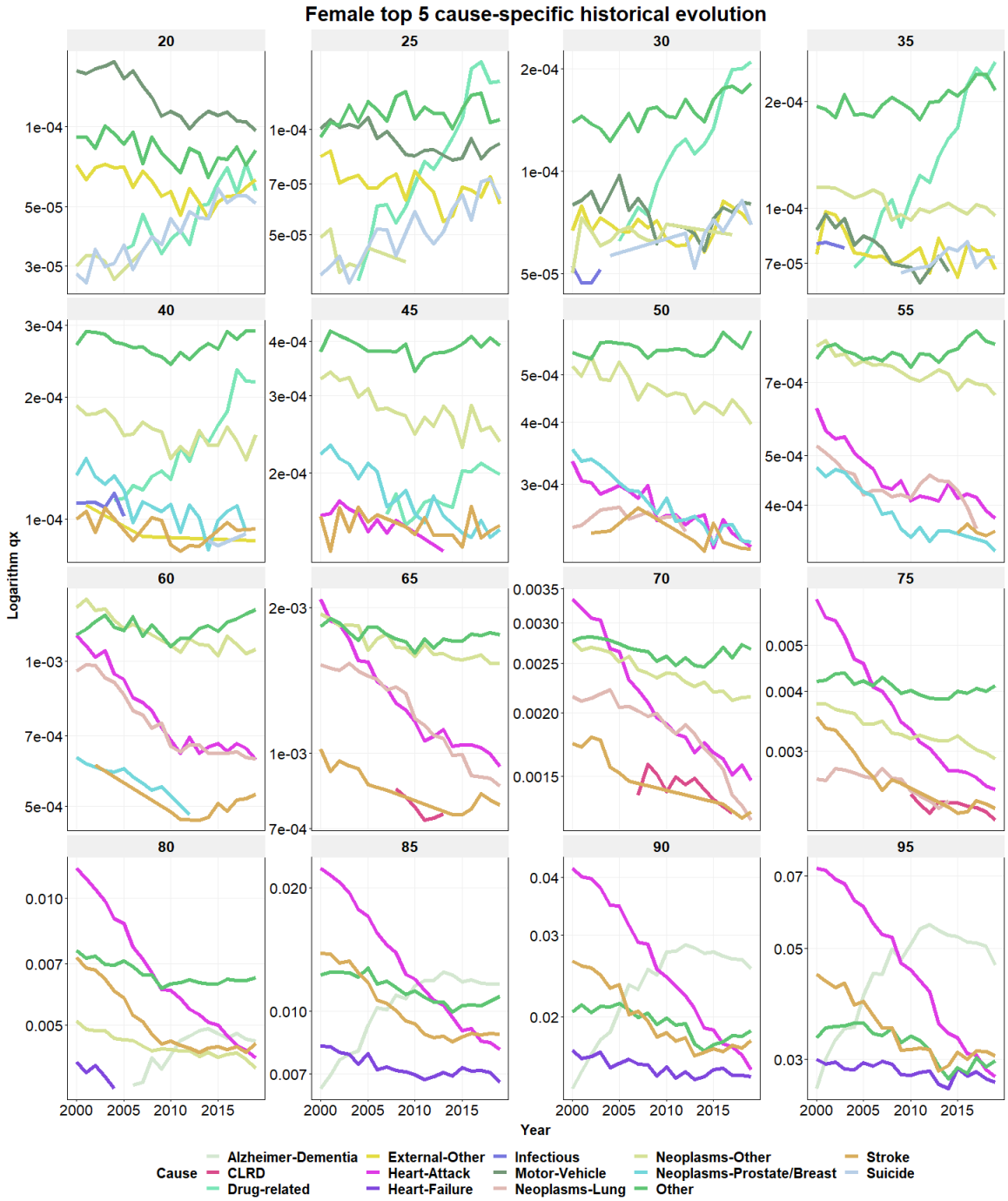


Figure 3.4: Top 5 female cause-specific historical evolution

3.3 Modelling and forecast results

This section introduces the central trajectory prediction which is based on the independent assumption among causes. The modelling steps consist of:

- Fit independently Poisson log-bilinear model for each cause

- Under the above independence assumption, each cause is modelled independently and the aggregate mortality rates could be expressed as the sum of cause-specific mortality rates. Poisson log-bilinear model is applied to each cause to forecast its central trajectory.

$$\mu_{x,t,i} = exp(\alpha_{x,i} + \beta_{x,i}\kappa_{t,i})$$
$$D_{x,t,i} \sim Poisson(E_{x,t} * \mu_{x,t,i}) = \frac{(E_{x,t} * \mu_{x,t,i})^{D_{x,t,i}} e^{-E_{x,t}*\mu_{x,t,i}}}{D_{x,t,i}!}$$

  from which

$$\kappa_{t+1,i} = \kappa_{t,i} + \delta + \epsilon_{t,i}$$
$$\Sigma = cov(\epsilon_{t,i}, \epsilon_{t,j})$$

  $\kappa_{t,i}$ is further extrapolated for the forecast:

$$\widehat{\kappa_{t+1,i}} = \kappa_{t,i} + \delta$$
$$\widehat{\mu_{x,t+1,i}} = exp(\alpha_{x,i} + \beta_{x,i}\widehat{\kappa_{t+1,i}})$$
$$\widehat{\mu_{x,t+1}} = \sum_i \widehat{\mu_{x,t+1,i}}$$

- Model each $\kappa_{t,i}$ by ARIMA (0,1,0) and collect residuals $\epsilon_{t,i}$ of each cause.

Figure 3.5 and Figure 3.6 compare the trend parameters from each cause and from the All-Cause model where the bold line represents the trend factor from the All-cause model, *Drug-related*; *Alzheimer - Dementia*;*Alcohol-related* and *Suicided* are highlighted by solid line. It can be observed that they share the opposite trend as in the All-Cause model, *Suicide; Alcohol-related; Drug-related; Alzheimer-Dementia* all have increased between 2000-2019. Especially *Drug-related* cause proves a significant lift compared to other causes. *Alzheimer-Dementia* also has experienced a different trajectory but the increase tends to slow down starting from 2011. Appendix 4.5.1 shows the two other parameters of $\alpha_{x,i}$ and $\beta_{x,i}$ of each cause.

As mentioned in subsection 2.2.2, the model forecast of the Poisson log-bilinear model is realized by extrapolation of $\kappa_{t,i}$. Therefore the future mortality force trajectory is highly impacted by the drift parameter δ fitted in the historical period, this drift is assumed to continue during the whole forecast horizon. The mortality risk is often involved in long-term uncertainty, in the meantime in order to capture the most recent experience which carries more valuable information, especially as observed in the historical period, most causes showed a turning point around 2010. As a first tentative and to test model suitability, the forecast horizon is set to 20 years, from 2020 to 2039.

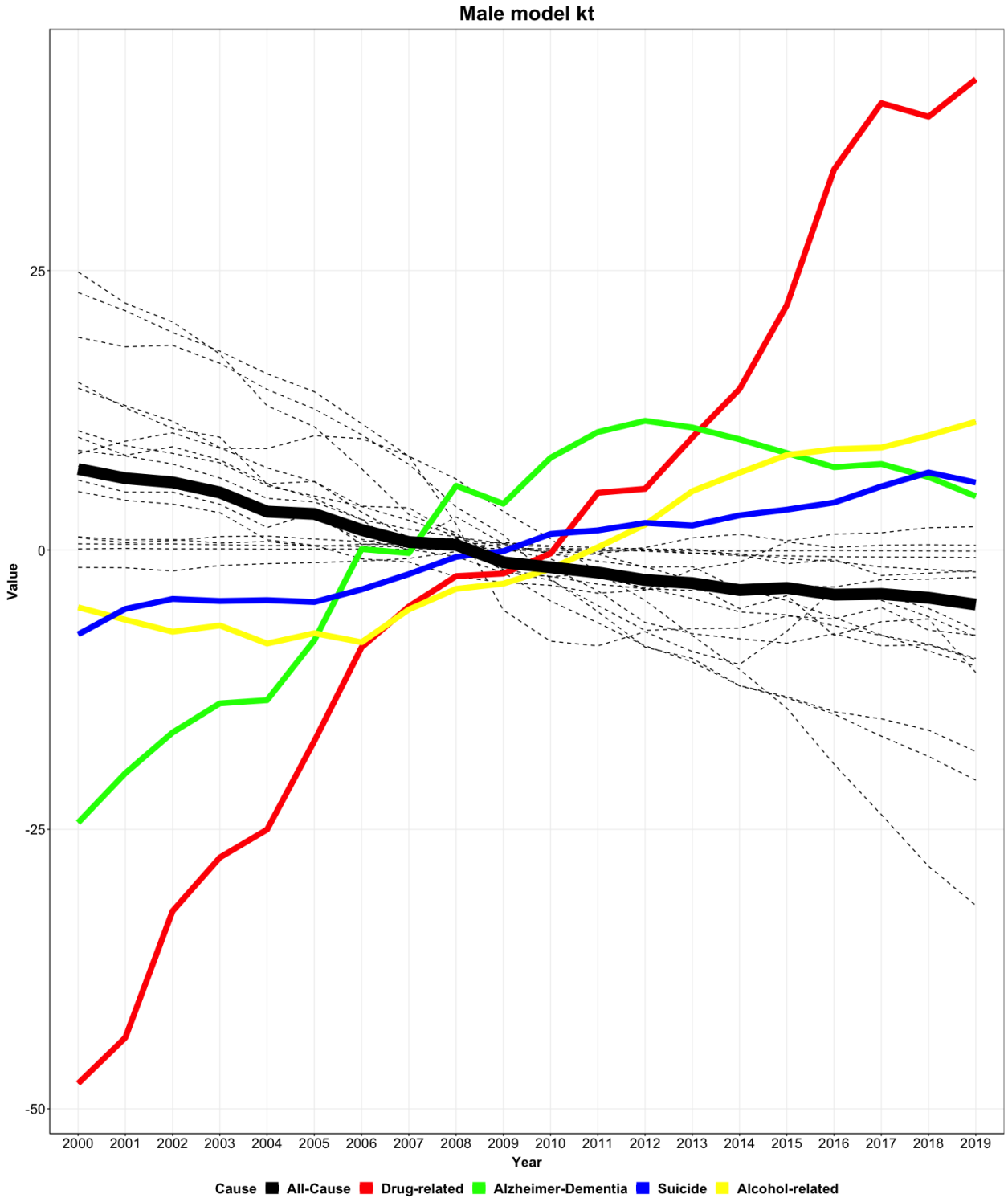


Figure 3.5: All-Cause and By-Cause male model κ_t

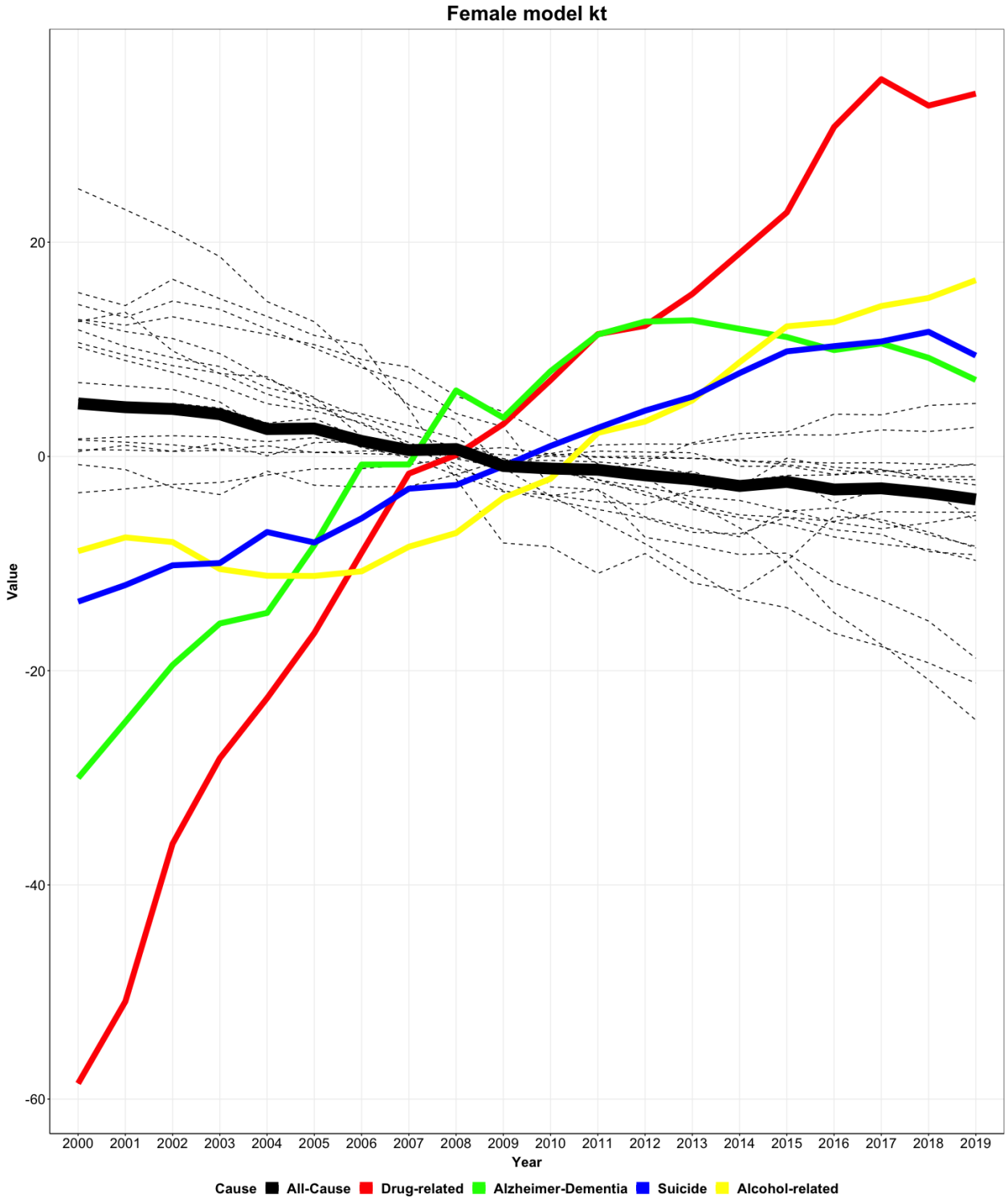


Figure 3.6: All-Cause and By-Cause female model κ_t

For the purpose of comparing the By-Cause model with the independence assumption and All-Cause model output, Figure 3.7 and Figure 3.8 show the sum of all causes forecast mortality rates which constitute the By-Cause aggregate mortality forecast. Some conclusions could be drawn from the comparison below:

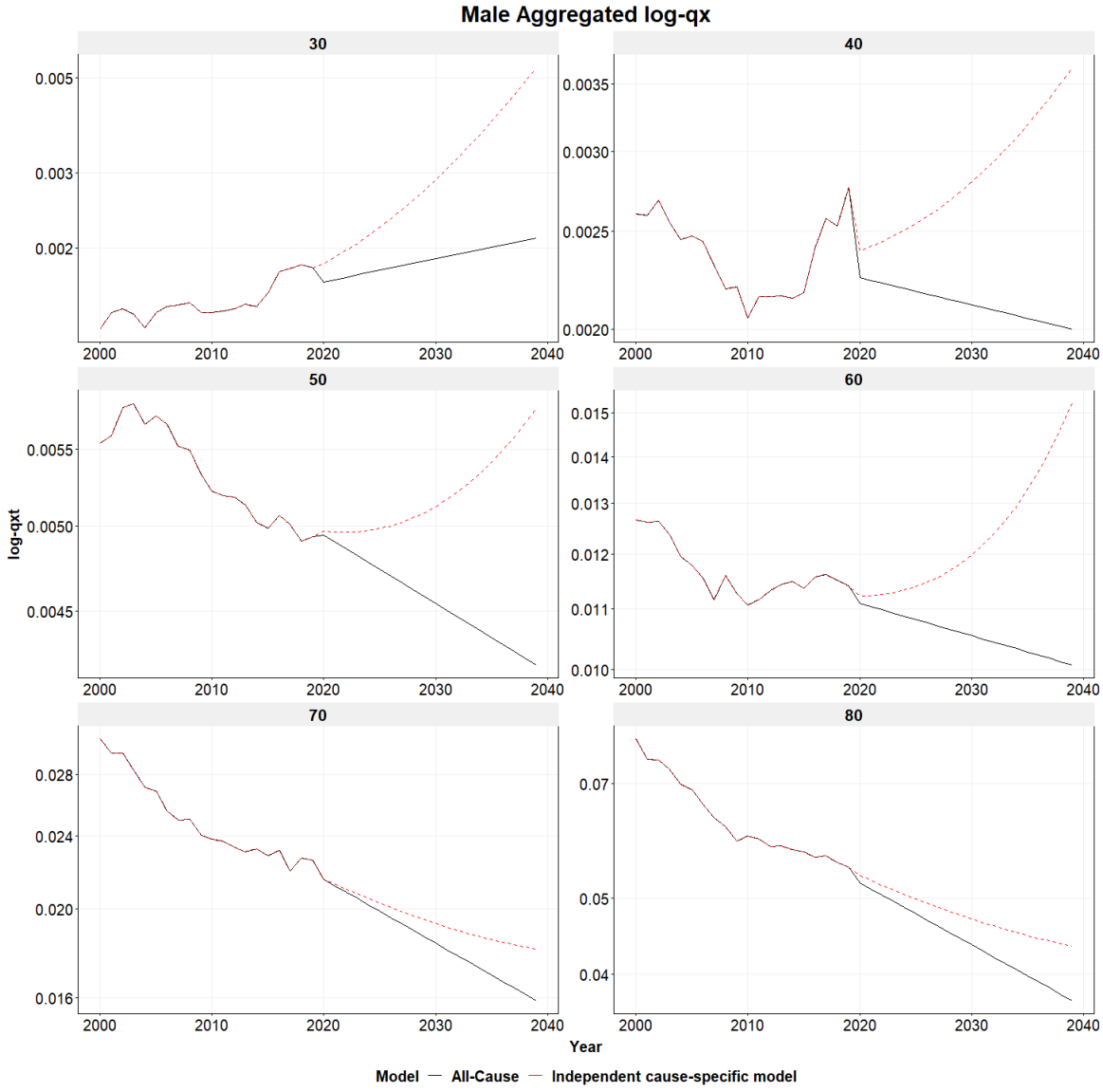


Figure 3.7: All-Cause and By-Cause male model forecast

- According to forecast results, the independence assumption model presents more pessimism than the All-Cause model, this effect is mainly driven by *Drug-related* and *Alzheimer-Dementia* cause

- Trend differences between All-Cause and By- Cause for ages between 30 and 59 is mainly driven by *drug and other*, which registered a highly increasing trend in the past years, especially 30-54. Furthermore, for certain ages 30-35, the mortality rate has already shown no sign of improvement in the past years.

- The stagnation of the Cardiovascular-family cause improvement and the increase of *Alzheimer-Dementia* are the main reasons for mortality improvement slowdown in the By-Cause model for older ages. However the increase is relatively less than younger ages, which is in line with the recent slowdown of *Alzheimer-Dementia* after

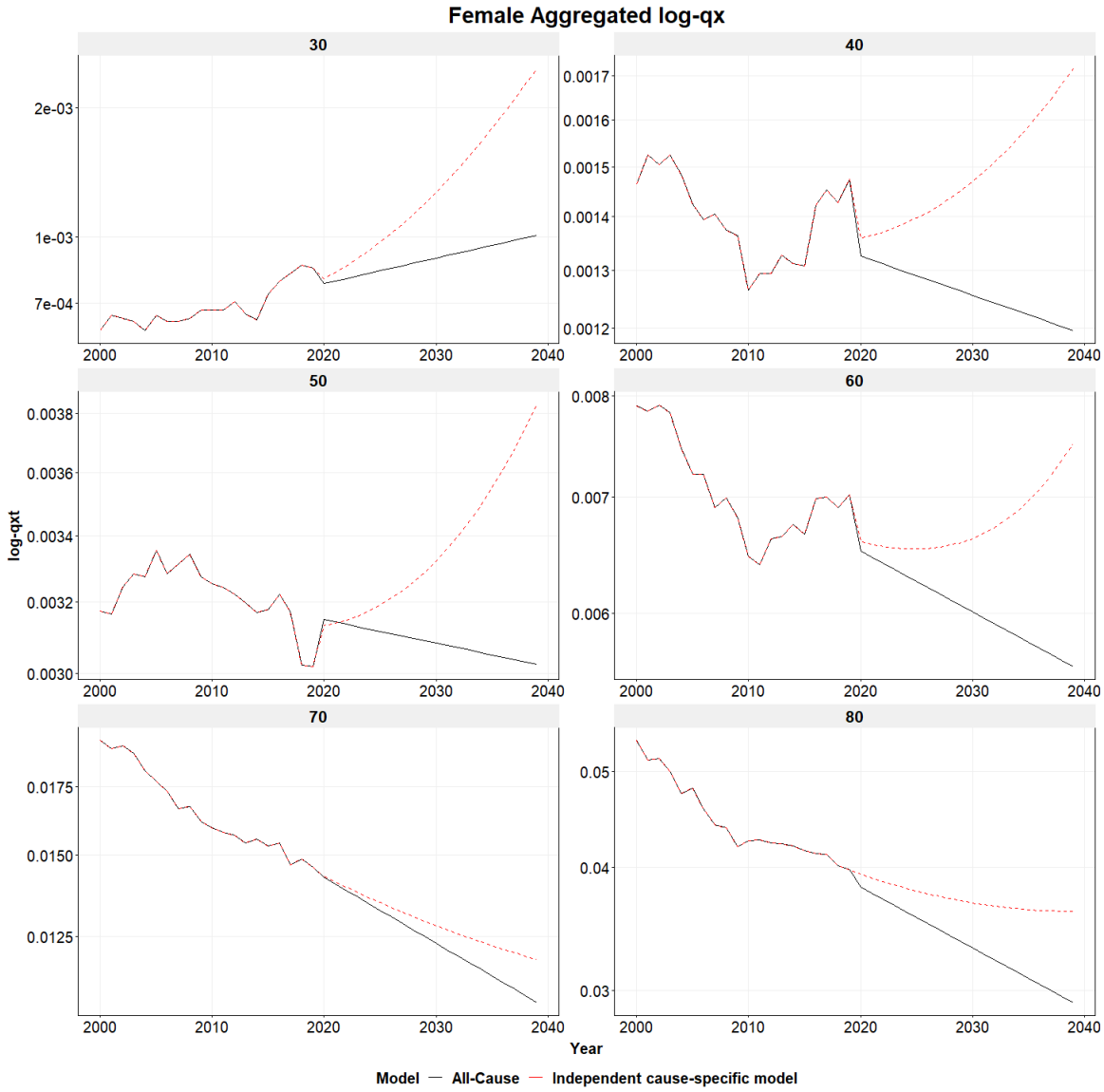


Figure 3.8: All-Cause and By-Cause female model forecast

2011. Under the linear extrapolation forecast method, *Alzheimer-Dementia* still has an important trend.

3.3.1 Prediction interval

The next step consists of building a prediction interval of aggregate mortality rates, aggregate level uncertainties depend on the dependence structure between causes in order to take into account mutual diversification between causes. This dependence structure is assumed to be captured by the correlation matrix between residuals $\epsilon_{t,i}$ of $\kappa_{t,i}$.

Figure 3.10 and Figure 3.9 display the correlation matrix of residuals $\epsilon_{t,i}$ of $\kappa_{t,i}$ for male and female. The correlation matrix contains pairwise correlation coefficients between causes and is used to simulate scenarios.

While the correlation matrix is automatically derived from the data and underlying model, the resulting matrices do not have a stable structure, it is highly sensitive to the choice of calibration period and the number of data points.

Based on the variance-covariance matrix Σ from residuals of $\kappa_{t,i}$, the future trajectory of each $\kappa_{t,i}$ can be simulated by taking into account dependence structure within causes by assuming a multivariate normal distribution between residuals of $\kappa_{t,i}$ and through 10000

simulations:

- Cholesky decomposition of variance-covariance matrix $\Sigma = RR^T$
- Generate independently and identically distributed standard normal distribution vector z
- The target simulated multivariate vector could then be written as ϵ = R*z
- Add simulated vector ϵ to initial central trajectory vector $\kappa_{t,i}$ recalculate the forecast of each cause and corresponding new aggregate mortality probabilities forecast.

Figure 3.11 and Figure 3.12 showed the resulting prediction interval for each gender, it can be seen that male and female both generate more uncertainties towards the end of the forecast horizon.

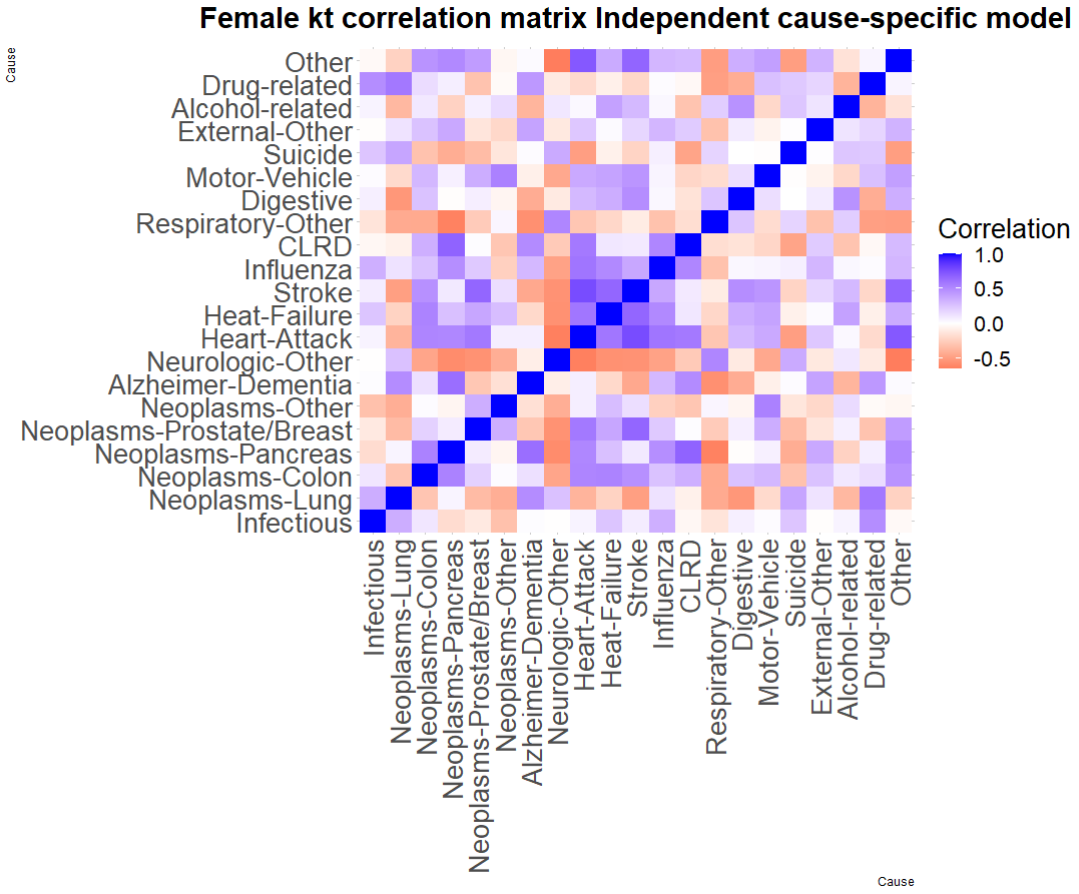


Figure 3.9: Female correlation matrix

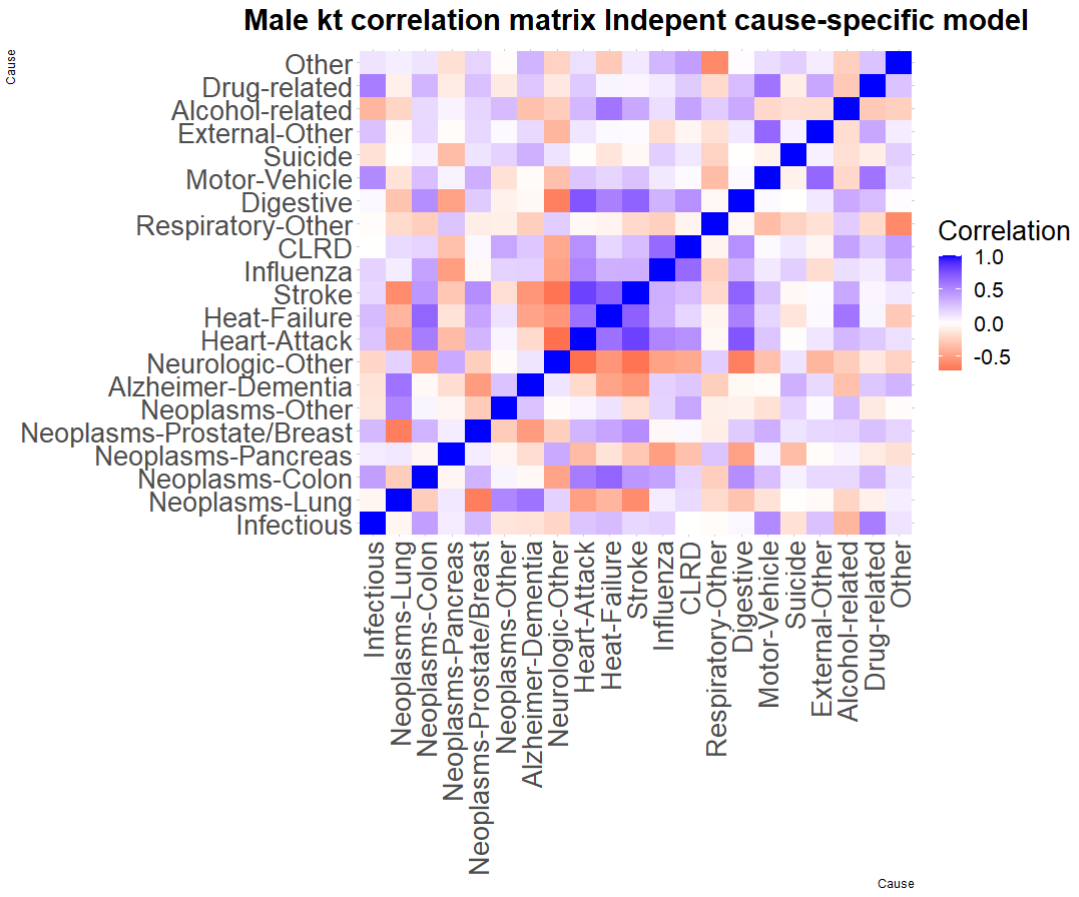


Figure 3.10: Male correlation matrix

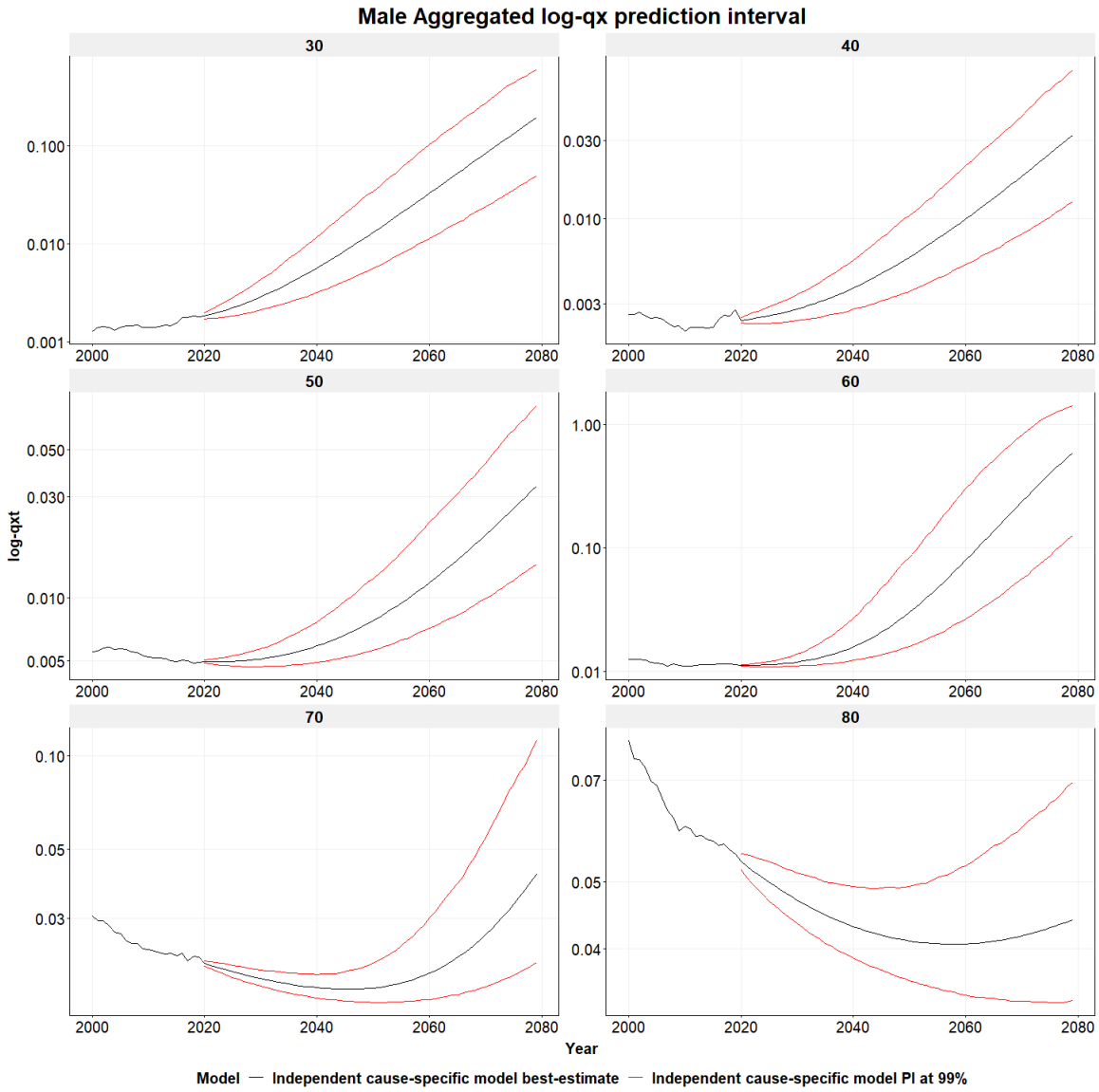


Figure 3.11: Male prediction interval between [0.5% , 99.5%]

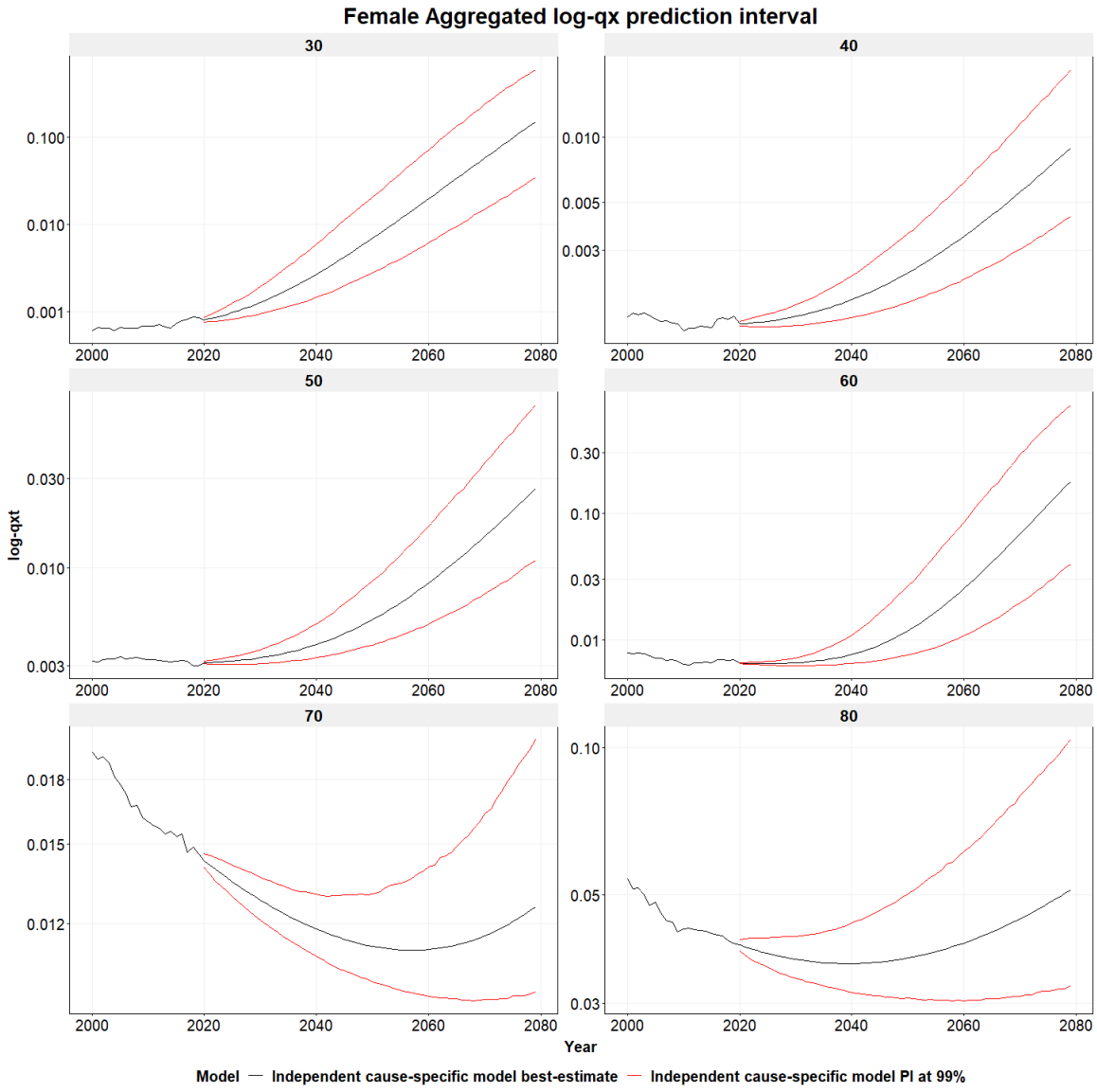


Figure 3.12: Female prediction interval at [0.5% , 99.5%]

3.4 Life expectancy

Life expectancy is a crucial measure of the average (residual) lifetime of an individual, as a function of death rate (survival probability), investigating such a concept could offer a more comprehensive view of risk encountered by the insurance company, not only for the mortality risk factor. In this section, life expectancy has been calculated using the following assumptions and inputs as described in Table 3.1:

	All-Cause	By-Cause
Age	0-110	By-Cause aggregate mortality rates forecast from age range 20 to 95 + All-Cause Best-estimate aggregate mortality rates for 0-19 and 96-110
Method	Period life expectancy	Period life expectancy

Table 3.1: Inputs and assumption

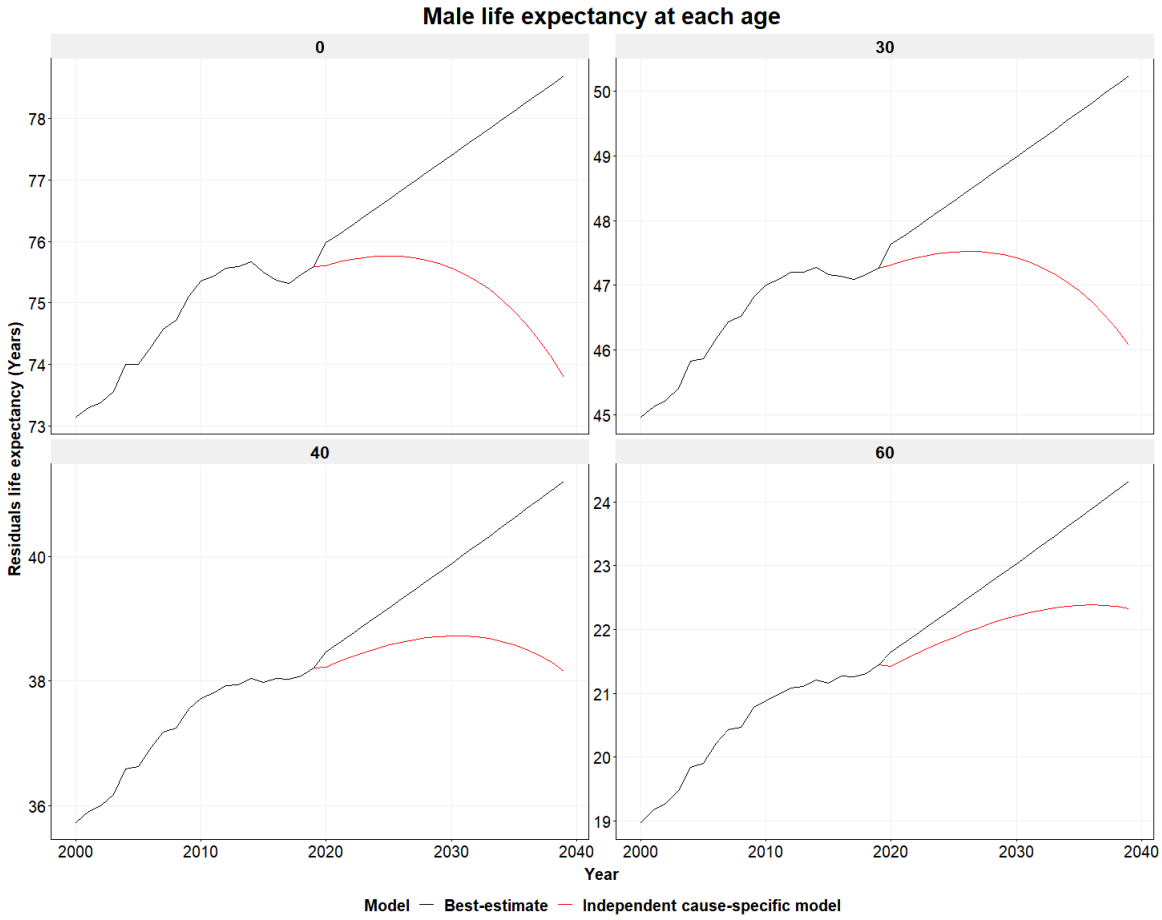


Figure 3.13: All-Cause and By-Cause male residual life expectancy at different ages

Since only ages between 20-95 are modelled in By-Cause model, the All-Cause best-estimate mortality forces of 0-19 and 96-110 are used to complete the table, thus the difference displayed in the life expectancy is only due to the 20-95 projection in By-Cause and All-Cause models. Furthermore, in this section life expectancy is calculated using the period life table, the period table calculates the mortality rates from a single year and assumes that for the rest of life, the mortality rates will remain the same.

Figure 3.13 and Figure 3.14 illustrate the period life expectancy from All-Cause and independent cause-specific model. As per the results of the period life expectancy forecast, the By-Cause model produced a more pessimistic result . This can be observed in the

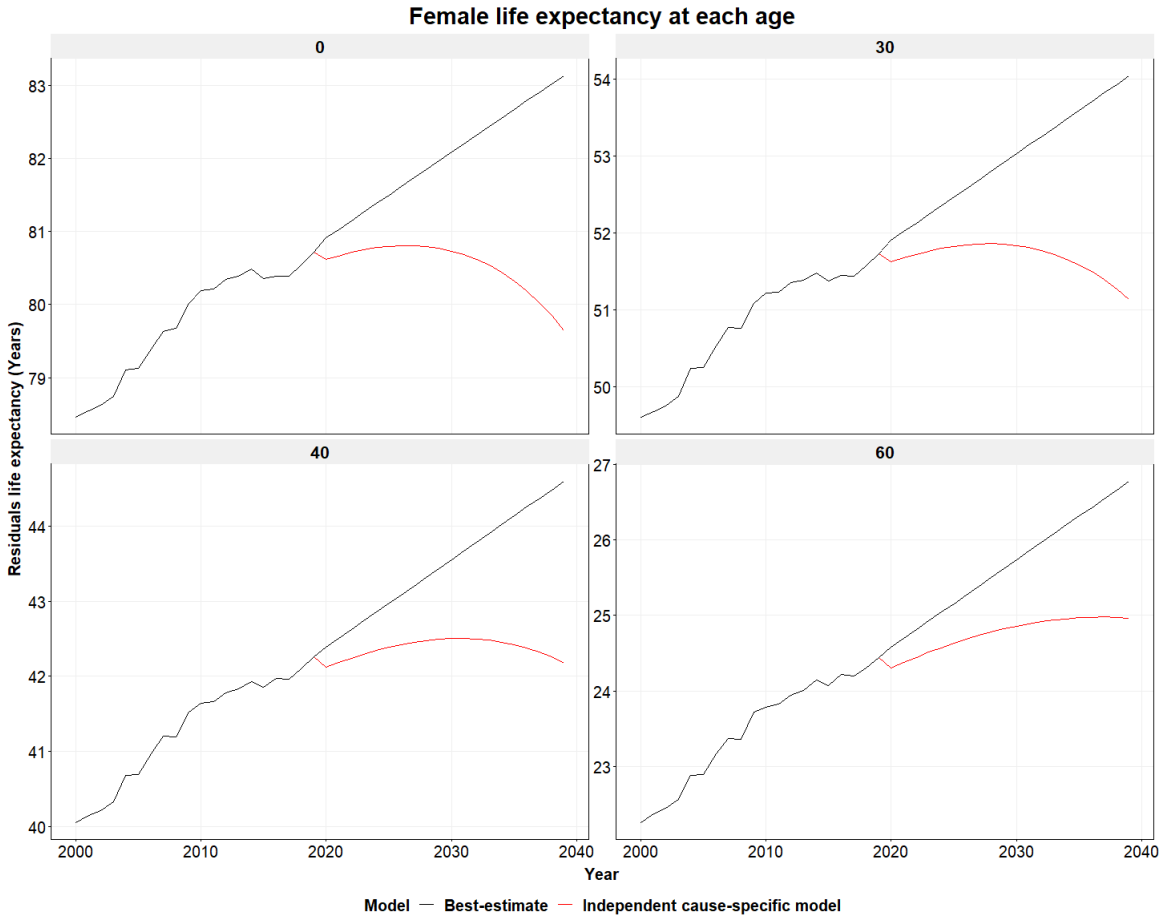


Figure 3.14: All-Cause and By-Cause female residual life expectancy at different ages

life expectancy at birth which takes into account all major causes' impact throughout the whole life, which demonstrates even unrealistic results as the decline of almost 2 years of expected lifetime in 20 years. This pessimism is mainly conducted by the underlying assumption in the modelling approach, each cause will independently follow the historical trend and continue until the end of forecast horizon, therefore *Drug-related* that has shown a rapid increasing trend in the historical period is assumed to follow the same increasing trend in the forecast horizon. *Drug-related* impacts mainly young and middle age groups, which explains the reason why there is more pessimism or even unrealistic result in life expectancy at birth than at age 60.

3.5 Limits

While an independent cause-specific model provides advantages in modelling and sheds light on individual characteristics of cause evolution, it may be deemed unrealistic when the forecast horizon is expanded to a longer period. Even at a 20-year forecast, life expectancy at birth already displayed a reduction of 2 years of life in 20 years, which is a relatively implausible scenario for a best-estimate forecast.

As mortality risk needs to be projected long term in (re-) insurance companies, a longer period forecast is calculated using the model. Figure 3.15 and Figure 3.16 display the forecast of aggregate mortality rates in 60 years. Resulting aggregate mortality rates from the independent cause-specific model are at least doubled compared to the All-Cause model, after 60 years in 2080.

It should be noted that the calibration period was set to 20 years in order to ensure

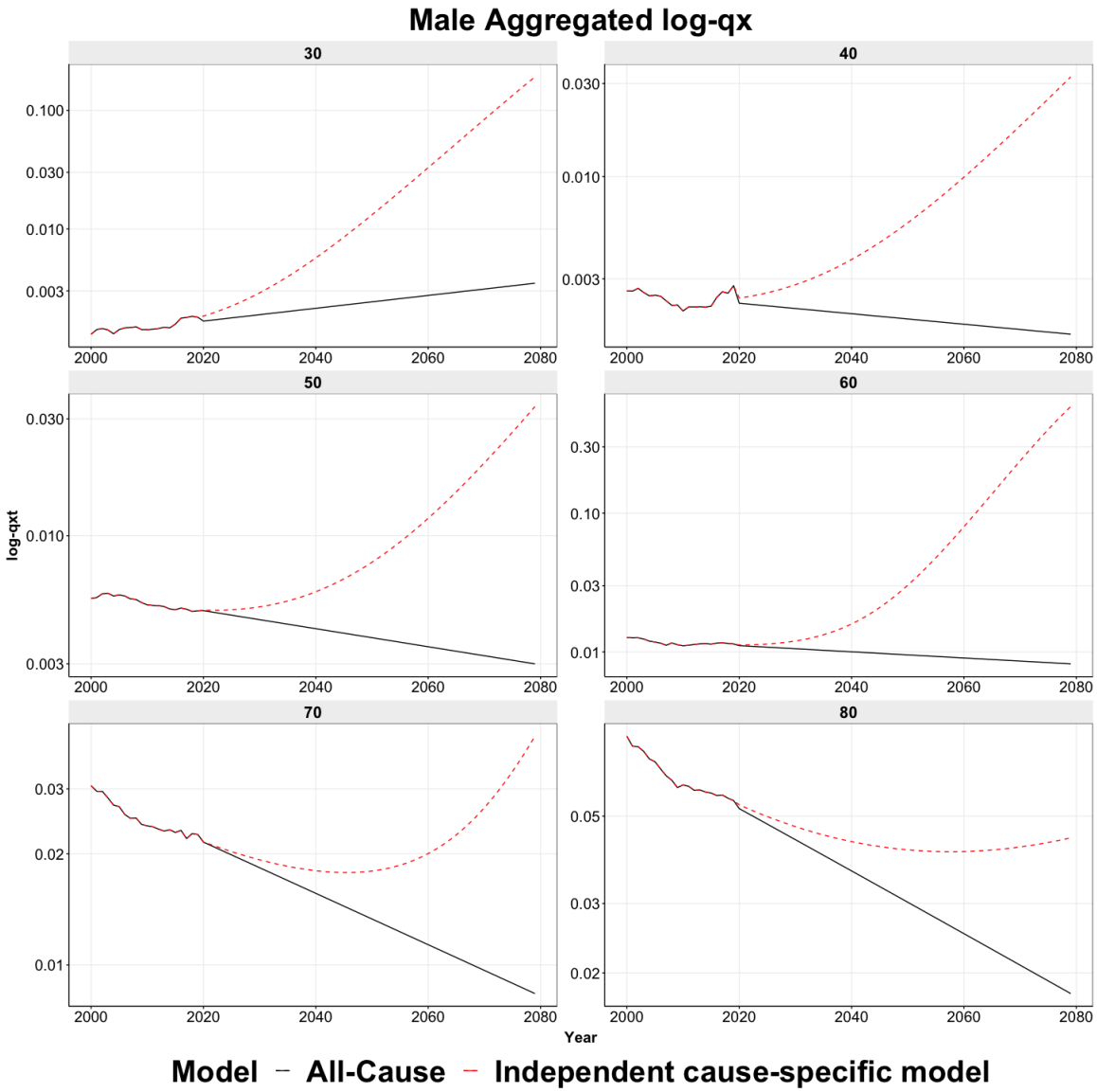


Figure 3.15: All-Cause and By-Cause male mortality probabilities forecast for 60 years

cause-of-death data stability, a forecast horizon of 60 years may exaggerate the historical trend of each cause observed in the last 20 years.

Furthermore, under the assumption of independent structure and linear extrapolation method, certain increasing causes may easily become dominant in the future, Figure 3.17 illustrates the aggregate and *Drug-related* log-mortality rates forecast for male aged 30, *Drug-related* that has demonstrated a sharp increase in the calibration period will become dominant in the forecast in the young ages.

Another potential explication is already cited in Wilmoth [1995] as dis-aggregated forecasts always output more pessimistic results, thus lack of dependence structure on best-estimate projection may not be suitable in mortality risk evaluation.

As a potential solution, the unrealistic forecast of the long-term forecast of the independent cause-specific model could be adjusted by applying expert judgement, for example by using breakpoint method to select most recent trend; setting up upper boundary of the end year of forecast of *Drug-related* mortality rates and adjusting the drift in the κ_t of *Drug-related*; or directly by fixing its mortality improvement a positive value.

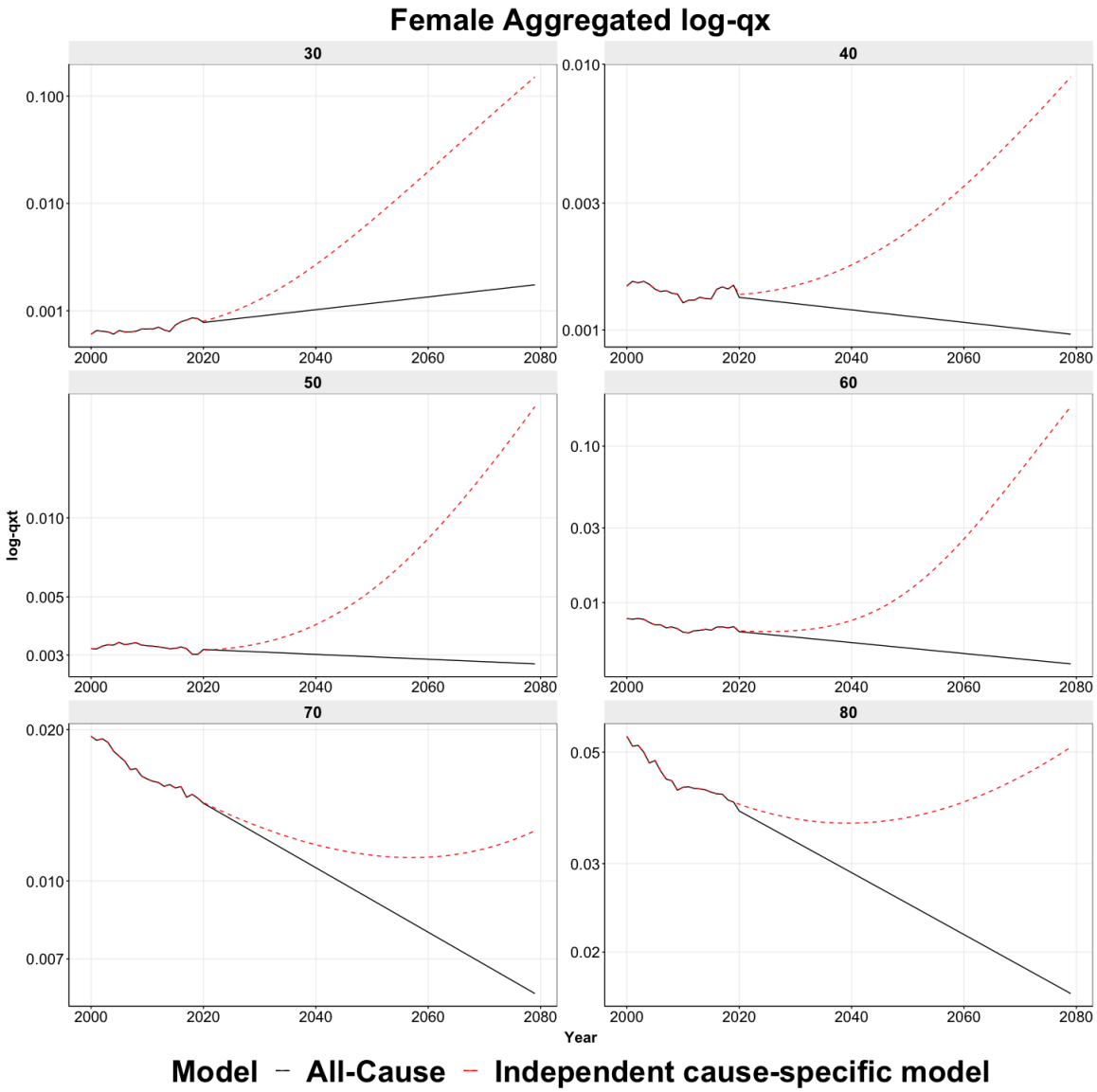


Figure 3.16: All-Cause and By-Cause female mortality probabilities forecast for 60 years

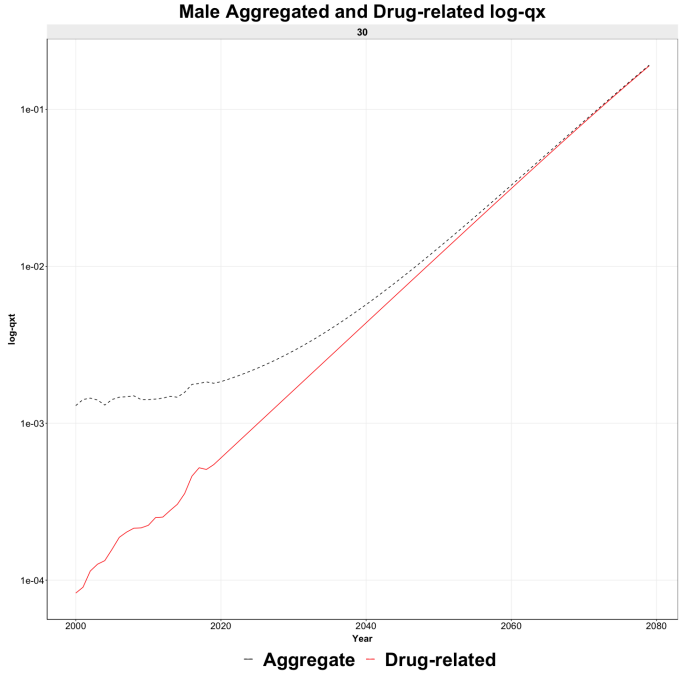


Figure 3.17: Aggregated and *Drug-related* log-mortality rates forecast for male aged 30

Chapter 4

Compositional data analysis

Independent assumption presented in the previous chapter, due to its simplicity, long-term forecast based on this model may produce unrealistic results at the aggregate level, which suggests it might not be suitable for internal use. Among the alternative cause-of-death modelling approaches which could address these issues above, Compositional data analysis suggests imposing an aggregate level constraint and forecasting each cause as proportion to ensure a coherent forecast. It is realized by its intrinsic mechanism to incorporate dependence structure between causes, the increase of proportion from a cause must induce the decrease of others under the sum constraint. This chapter intends to explore Compositional data analysis for the purpose of coherent cause-specific forecasts with respect to aggregate level evolution.

Two sub-models of Compositional data analysis are tested which differ in the trend parameter choice, one proposes to assume every cause follows a shared trend and the other suggests allocating each cause a trend factor in order to capture cause-specific dynamics.

Table 4.1 summarizes the difference between the three models tested and their respective advantages and limits.

Table 4.1: Model comparison

Model name	**Modelling approach**	**Advantages**	**Limits**
Independent cause-specific model	Model each cause-of-death independently	Easy implementation and emphasize on individual cause characteristics	Unrealistic long-term forecast of aggregate mortality rates
CoDa Common Trend	- Pre-determine aggregate mortality rates forecast as constraint - Model each cause's proportion - Assume a common trend for every cause	- Coherence of cause-specific forecast with respect to aggregate mortality rates. - Explanatory ability - Risk transfer between cause-of-death and ages	- Long term forecast predicting dominance (over 70%) of *Drug-related* cause - Common Trend for each cause
CoDa Multi Trend	- Pre-determine aggregate mortality rates forecast as constraint - Model each cause's proportion - Assume an individual trend for every cause	- Coherence of cause-specific forecast with respect to aggregate mortality rates. - Explanatory ability - Specific trend evolution for each cause - Risk transfer between cause-of-death and ages	Long term forecast predicting dominance (over 70%) of *Drug-related* cause

4.1 Theoretical background of Compositional data analysis

Compositional data analysis (CoDa) [Aitchison [1982]] intends to apply common statistical techniques on non-negative values whose sum is constant, compositions could be seen as percentages or proportions of a whole, vary from 0 to a defined sum constraint. More importantly, they belong to a positive simplex space defined as below:

$$S^d = x_1, ..., x_d : x_i > 0, (\forall i = 1, ..., d), x_1 + ... + x_d < 1$$

Usual multivariate analysis isn't applicable in the compositional context, due to the intrinsic sum constraint barrier. Aitchison [1982] proposed a logistic transformation on initial compositional vector from S^d to R^d from a function: $f : S^d \to R^d$ in order to study compositional vector properties in an unconstrained space.

Several properties in the simplex space ought to be pointed out:

- For a d-dimension compositional vector X which satisfy sum constraint $\sum_i^d x_i = M, (\forall i = 1, ..., d)$
- Operator C: return transformed values $S = f(x)$ back to proportions. C(Y) = $[\frac{Y_i}{\sum Y_i}, ..., \frac{Y_j}{\sum Y_i}] \times M$
- $\oplus$: Association between two compositional vectors called Perturbation: $Z = X_1 \oplus X_2 = C(x_1^1 x_2^1, ..., x_1^d x_2^d)$
- $\ominus$: Perturbation with inverse element of another compositional vector: $X_1 = Z \ominus X_2 = C(z^1/x_2^1, ..., z^d/x_2^d)$

One of the feasible transformations proposed is the centred log-ratio *(clr-)* transformation, expressed as below with g: the geometric mean of the compositional vector

$$g = (x_1 * x_2 * .... * x_d)^{1/d}$$
$$U = clr(X) = [ln(\frac{x_1}{g}), ..., ln(\frac{x_d}{g})]$$
$$X = clr^{-1}(U) = C(e^{U_1}, ..., e^{U_d})$$

where U represents the transformed vector in real space and clr^{-1} is the inverse transformation returning transformed values into compositional data. Compared to other transformation functions, *clr*-transformation provides better interpretability since the greater the proportion, the greater the *clr*-transformed values. Because the transformed values belong afterwards to the real space, it is appropriate to implement statistical analysis over these values and transform them back into compositional vectors.

4.2 Cause-of-death modelling with CoDa

This section introduces the adaptation of CoDa model in the mortality risk modelling context.

4.2.1 Compositional Lee-Carter model

The first application of CoDa on cause-of-death modelling is developed by Oeppen et al. [2008], and two variants are proposed from Kjærgaard et al. [2019]. They suggest to use the CoDa Lee Carter model to produce a coherent forecast of life table death distribution under the constraint of life table radix. Deaths are redistributed from an age-cause group to other age-cause groups, one survived from age x_1 and cause i and died of cause j at age x_2, will reduce the mortality rates at the age-cause group (x_1, i), and increase the mortality rates of another age-cause group (x_2, j). This leads to the following clr-transformation and modelling approach:

$$clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i}\kappa_t + \epsilon_{x,t,i} \tag{4.1}$$

where $d_{x,t,i}$ represents the life table death of cause i, at age x and year t, $\alpha_{x,i}$ geometric mean of cause and age over year, $\beta_{x,i}$ measures the age-and-cause-specific sensitivity to trend factor κ_t, it describes the gain (loss) of deaths for an age and a cause in relative terms. For example, a positive $\beta_{x,i}$ value combined with a positive κ_t output a positive clr value, which means this causes i at this age x gains more proportions.

From this basis, Kjærgaard et al. [2019] proposes to allocate for each cause a trend parameter in order to capture each cause's own evolution:

$$clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i}\kappa_{t,i} + \epsilon_{x,t,i} \tag{4.2}$$

Another approach developed by Piveteau and Tomas [2018] suggested imposing an aggregate mortality rates forecast constraint by age and year, which is derived from an All-Cause model such as Poisson log-bilinear model, and links the cause-of-death distribution of each age to this aggregate mortality rates constraint. This approach allows projecting directly cause-specific mortality rates, and it carries out more explanatory values on cause-specific proportions related to classical common mortality modelling practice. Clr-transformation is done in this case within each fixed age $clr(d^x_{t,i} \ominus \alpha^x_i)$ and then stacked together to be decomposed under the hypothesis of Common Trend (see 4.2.2).

This latter approach has been retained for this thesis as alternative approach to model cause specific mortality rates, due to its closeness with the classical modelling approach, and its explanatory ability.

Furthermore this thesis also attempts to apply the Multi Trend variant suggested in Kjærgaard et al. [2019] under the latter approach's aggregate mortality rates constraint.

4.2.2 Modelling steps

The main idea is to apply Lee-Carter model on clr-transformed compositional vectors, and use the same extrapolation techniques on κ_t. The steps of CoDa modelling could be summarized as follow:

- Determine first aggregate mortality rates central trajectory forecast $q_{x,t}$ by a classical model such Lee-Carter.

- Restrain sum of the proportion of each cause to 1, the variable of interest is the distribution of each cause $\sum_i D_{x,t,i} = D_{x,t} => \sum_i q_{x,t,i} = q_{x,t} => (\sum_i q_{x,t,i})/q_{x,t} = 1$

- Obtain the matrix of mortality proportion $S = s_{x,t,i} = q_{x,t,i}/q_{x,t}$ with $q_{x,t}$ as All-Cause mortality rates

- Transform mortality proportion matrix S by centered log-ratio clr for each independent age group x: $clr(d^x_{t,i} \ominus \alpha^x_i)$ which is perturbed by the geometric mean of the age-composition at time t: g, stack all N age groups' clr-transformed values to a single matrix of dimension (NK $\times$ T): Y with N ages; K causes and T years
- Apply SVD on clr-transformed values matrix Y to obtain beta, kappa and subsequent forecast of clr-transformed values matrix Y^{Pred}
- Perform inverse transformation clr^{-1} of forecast clr-transformed matrix $Y^{Forecast}$ by the operator C(exp(Y)). The dependence between causes is realized by the operator C(exp(Y)), which normalizes the clr values into proportion.
- Derive cause-specific proportion forecast $S^{Forecast}$ which is multiplied by forecast $q_{x,t}$ determined in the first step.

The clr-transformed matrix Y could differ by the choice of model. All causes' proportions by age and year are stacked in a single matrix, for N ages, K causes and T years, the clr-transformed matrix Y is of dimension (NK $\times$ T). The single matrix leads consequently to an estimation of a unique κ_t, and as the underyling hypotheses is that all causes follow the same trend, the model is referred as CoDa Common Trend (CT).

	T_1	T_2	...	T_T
N_1, K_1	clr	clr	clr	clr
N_1, K_2	clr	clr	clr	clr
N_1, K_3	clr	clr	clr	clr
...	...	...	...	...
N_1, K_K	clr	clr	clr	clr
N_2, K_1	clr	clr	clr	clr
...	...	...	...	...
N_N, K_1	clr	clr	clr	clr
...	...	...	...	...
N_N, K_K	clr	clr	clr	clr

Table 4.2: Common Trend clr-transformed matrix

	T_1	T_2	...	T_T
N_1, K_1	clr	clr	clr	clr
N_2, K_1	clr	clr	clr	clr
...	...	...	...	...
N_N, K_1	clr	clr	clr	clr
$\vdots$				
	T_1	T_2	...	T_T
N_1, K_K	clr	clr	clr	clr
N_2, K_K	clr	clr	clr	clr
...	...	...	...	...
N_N, K_K	clr	clr	clr	clr

Table 4.3: Multi Trend clr-transformed matrices

Instead of assuming a common trend, another alternative is to take into account the distinct evolution of each cause by adjusting the stacked clr-transformed matrix Y as shown in Table 4.2.

Separate decomposition of K cause-specific transformed matrices, which leads to K matrices of dimension N x T, allows obtaining K trend factor $\kappa_{t,i}$. This model is referred as the CoDa Multi Trend (MT). The structure of the matrix is shown in Table 4.3.

After the forecast of each cause's matrix, all the matrices are stacked together again with the same dimension as in CoDa Common Trend (NK x T).

The last step of normalization realized by operator C(exp(Y)) is performed on the stacked matrix of dimension NK x T. This final step ensures to capture the dependence and interactions between the causes of death and the ages through the time.

4.3 Modelling and results

This section introduces the application of CoDa Common Trend (CT) and CoDa Multi Trend (MT) on CDC data. Age between 20-95 and years from 2000-2019 was selected to

ensure data stability towards classification standard changes, and sufficient cause-specific exposure among different ages.

4.3.1 Results

Clr-transformation

The first step consists of transforming initial cause-specific proportions into real space values by clr, Table 4.4 to Table 4.5 shows an example from female cause-specific proportions at age 20 to clr-transformed matrix. The clr transformed values are in line with the initial proportion order as the causes that weight more at age 20 present larger clr values as well.

Initial proportions	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Infectious	0.022	0.021	0.021	0.028	0.016	0.018	0.024	0.027	0.020	0.020
Neoplasms-Lung	0.001	0.002	0.002	0.002	0.003	0.004	0.002	0.001	0.002	0.001
Alzheimer-Dementia	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.001	0.001
Neurologic-Other	0.004	0.001	0.001	0.001	0.003	0.002	0.001	0.004	0.001	0.003
Heart-Attack	0.004	0.008	0.006	0.004	0.003	0.004	0.003	0.007	0.003	0.003
Heat-Failure	0.041	0.033	0.026	0.025	0.029	0.029	0.039	0.032	0.036	0.041
Stroke	0.033	0.014	0.020	0.018	0.017	0.020	0.019	0.010	0.018	0.021
Motor-Vehicle	0.349	0.354	0.363	0.349	0.376	0.333	0.356	0.319	0.296	0.277
Suicide	0.060	0.058	0.076	0.061	0.066	0.079	0.064	0.077	0.088	0.091
External-Other	0.154	0.142	0.154	0.149	0.150	0.155	0.130	0.153	0.144	0.139
Alcohol-related	0.005	0.004	0.003	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Drug-related	0.039	0.047	0.039	0.050	0.041	0.075	0.079	0.105	0.090	0.084
Other	0.196	0.205	0.182	0.208	0.203	0.189	0.209	0.163	0.210	0.202

Table 4.4: Cause-specific proportion of Female at age 20

Clr	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Infectious	0.949	0.820	0.823	0.841	0.841	0.830	0.859	0.814	0.775	0.757
Neoplasms-Lung	−1.811	−1.751	−1.821	−1.853	−1.904	−1.931	−2.013	−2.023	−2.040	−2.050
Alzheimer-Dementia	−2.678	−2.680	−2.702	−2.710	−2.733	−2.738	−2.763	−2.764	−2.763	−2.753
Neurologic-Other	−1.949	−2.488	−2.285	−2.147	−2.032	−1.966	−1.664	−1.727	−1.737	−1.691
Heart-Attack	−0.693	−0.659	−0.711	−0.733	−0.769	−0.791	−0.849	−0.862	−0.881	−0.894
Heat-Failure	1.194	1.286	1.268	1.250	1.243	1.238	1.203	1.219	1.226	1.218
Stroke	0.728	0.820	0.756	0.722	0.684	0.657	0.578	0.571	0.553	0.530
Motor-Vehicle	3.695	3.732	3.678	3.654	3.630	3.594	3.538	3.506	3.461	3.416
Suicide	1.811	1.905	1.975	1.988	2.043	2.083	2.135	2.196	2.255	2.280
External-Other	2.790	2.845	2.812	2.794	2.782	2.762	2.723	2.712	2.690	2.662
Alcohol-related	−1.806	−1.457	−1.710	−1.842	−2.008	−2.106	−2.424	−2.436	−2.487	−2.545
Drug-related	1.457	1.535	1.677	1.717	1.820	1.897	2.021	2.118	2.219	2.275
Other	3.059	3.094	3.080	3.071	3.072	3.059	3.043	3.034	3.018	2.993

Table 4.5: Clr-transformation values of Female at age 20

Choice of rank

As explained above, a compositional Lee-Carter model is employed and parameters estimation is realized by Singular Value Decomposition (SVD), which requires low-rank approximation. Figure 4.1 and Figure 4.2 show the explained variance cumulative percentage from CoDa Common Trend and CoDa Multi Trend, in order to set a unique rank number for both models and more importantly for each cause in the CoDa Multi Trend. Rank 3 seems to be a reasonable choice of low-rank approximation since the 3 first ranks explain at least 90% of the total variance.

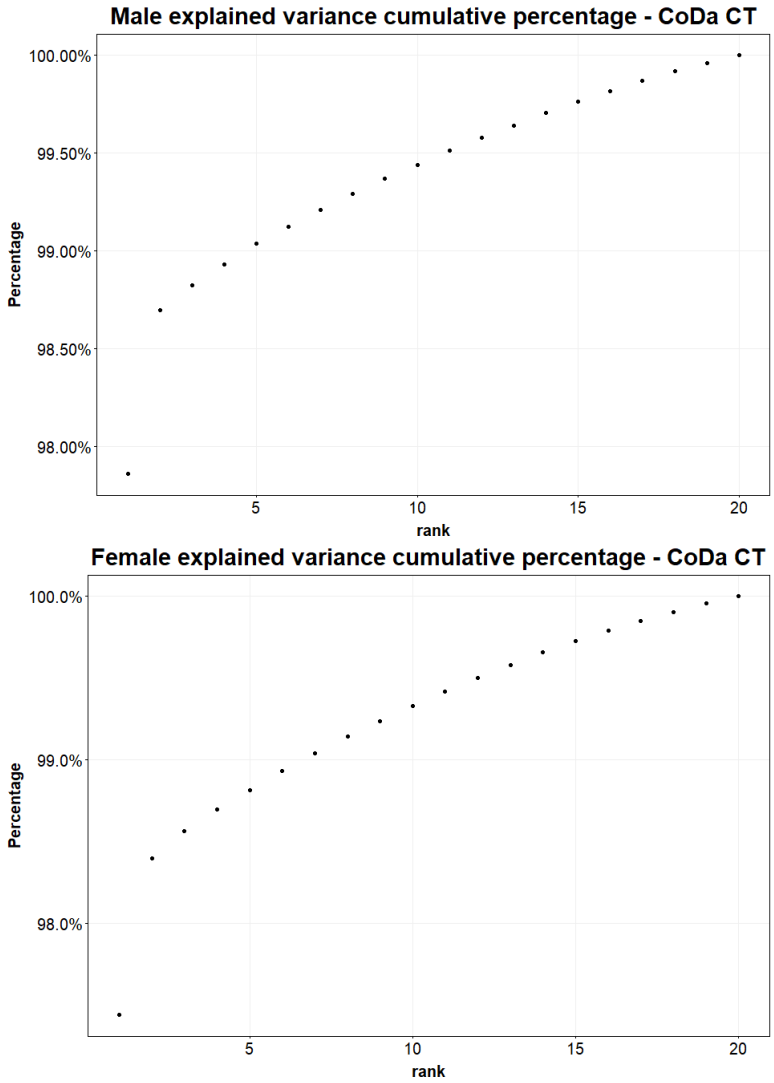


Figure 4.1: CoDa Common Trend explained variance percentage

Common Trend: parameters interpretation and risk transfer concept

Figure 4.3 and Figure 4.4 demonstrate the first 3 ranks of κ_t and the first rank of $\beta_{x,i}$ accounted for the fit and forecast of clr-transformed values matrix Y^{Pred} in CoDa Common Trend. It could be observed that the first rank of κ_t, which has the biggest value, is always negative for both male and female, the sign of κ_t should be always associated with the sign of $\beta_{x,i}$.

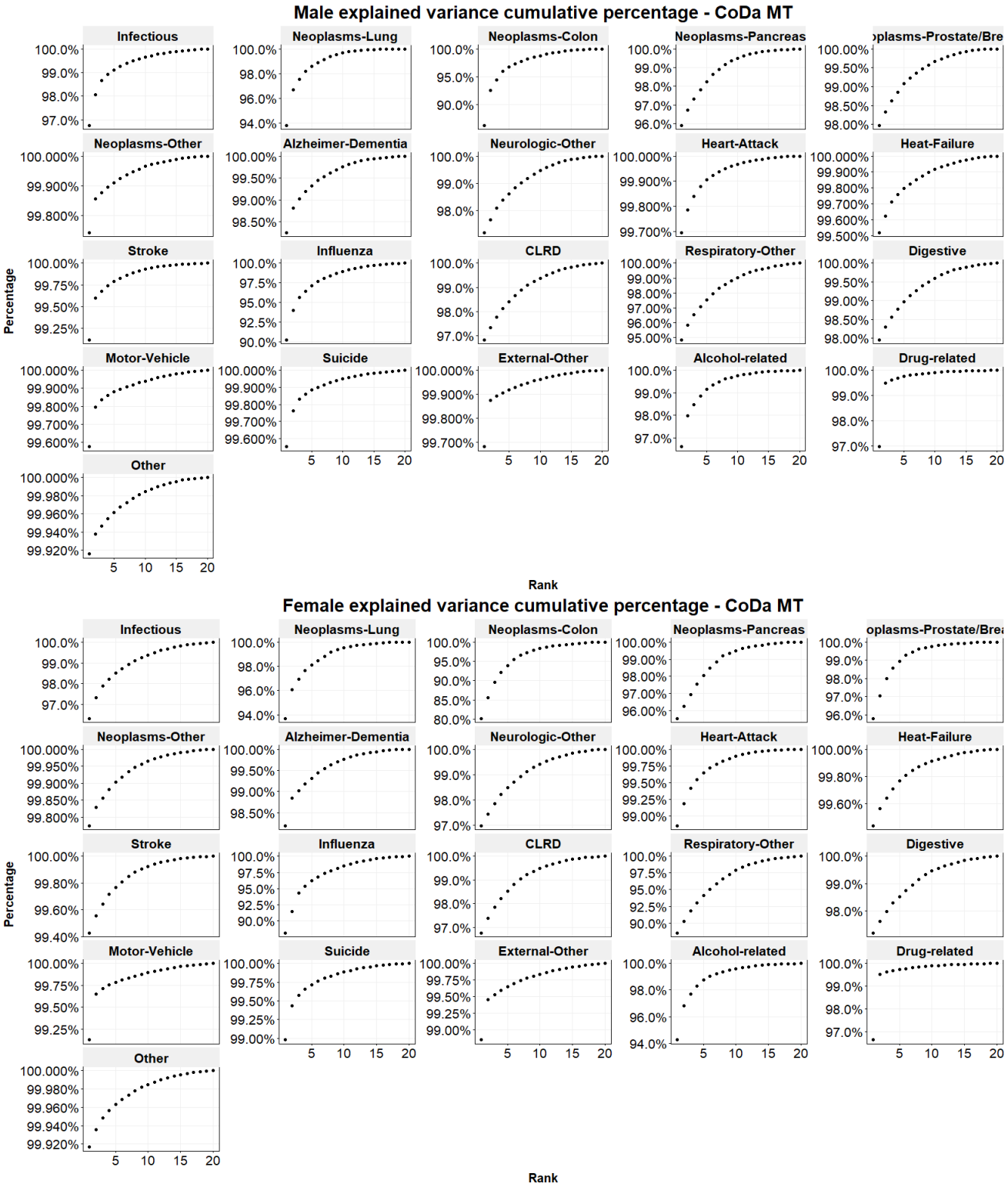


Figure 4.2: CoDa Multi Trend explained variance percentage

The combination of the sign of these two parameters indicates qualitatively the relative importance of each cause within a fixed age. Focusing on the first rank of the parameters and as an example for male and female aged 20, κ_t is negative, *Neoplasm - Prostate/Breast* has a very positive $\beta_{20,Neoplasm-Prostate/Breast}$ while *Drug-related* has a very negative $\beta_{20,Drug-related}$. This means that, combined with the value of κ_t , during the period, the cause *Drug-related* is gaining additional deaths while for the *Neoplasm - Prostate/Breast* cause is having less deaths. This also demonstrates the explanatory ability of CoDa models.

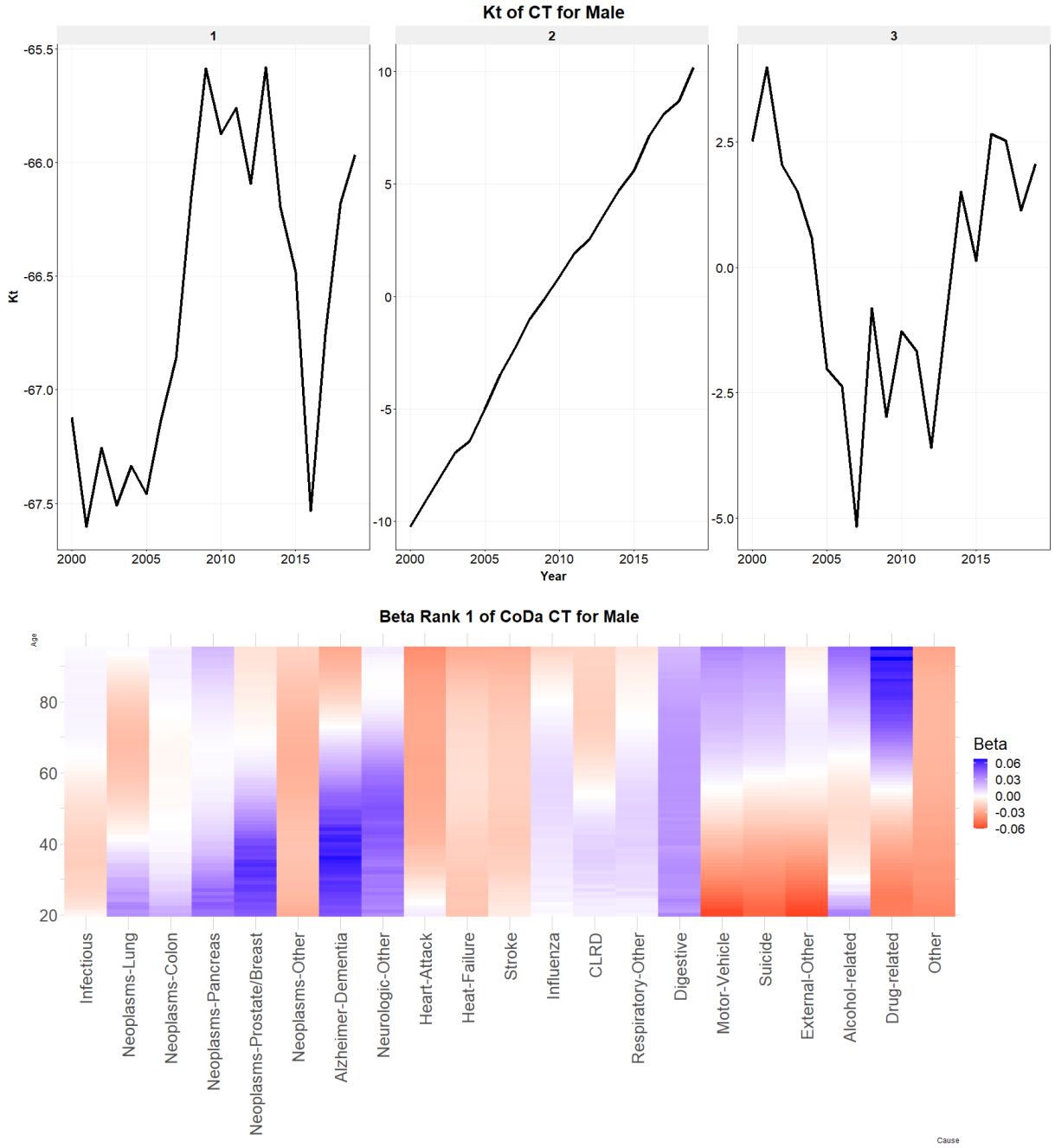


Figure 4.3: CoDa Common Trend κ_t and $\beta_{x,i}$ for male

Based on $\beta_{x,i}$ in Figure 4.3 and Figure 4.4, female and male maintain similarly the same cause-of-death distribution for each age, except for *Neoplasm-Prostate/Breast* which by nature has different target ages.

An important mechanism and adding value of CoDa framework is the concept of risk transfer, the reduction of the mortality risk in a cause or at an age, will result in an increase in other causes and ages. The mechanism of risk transfer can be better understood via the CoDa framework by the difference of $\beta_{x,i}$.

It is worth mentioning that the quantification of this mechanism of transfer of risk in the CoDa Common Trend is not straightforward to determine, because it is not only impacted by the parameters κ_t and $\beta_{x,i}$ of each cause, initial proportion and other causes' parameter are also needed to be considered in the quantification. The method to quantify

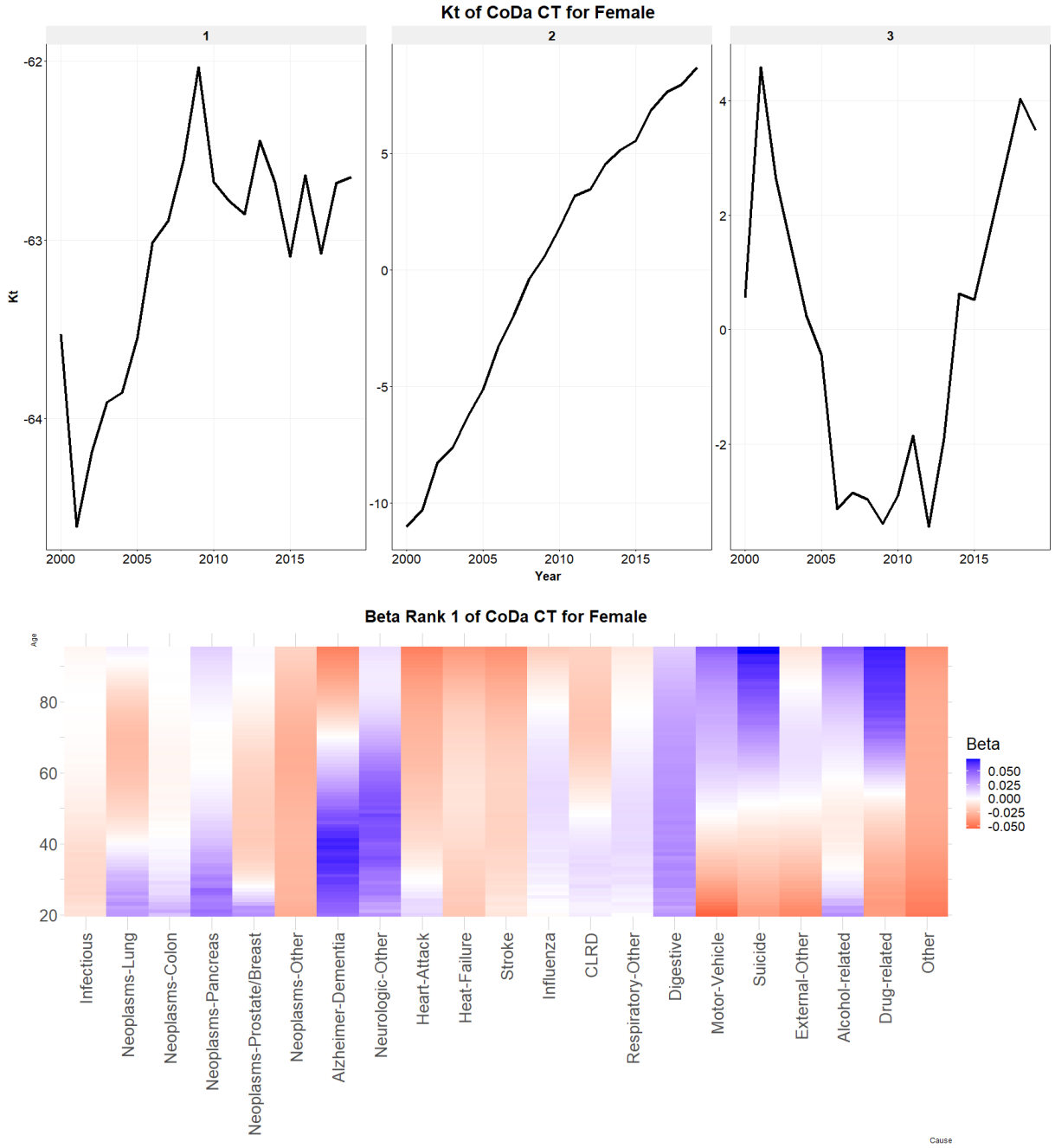


Figure 4.4: CoDa Common Trend κ_t and $\beta_{x,i}$ for female

the risk transfer is proposed in Piveteau and Tomas [2018] by computing:

$$\frac{\partial \frac{D_{x,t,i}}{D_{x,t}}}{\partial t} = \sum_i (\beta_{x,i} - \beta_{x,j}) \frac{\partial k_t}{\partial t} E[\frac{D_{x,t,i}}{D_{x,t}}] E[\frac{D_{x,t,j}}{D_{x,t}}] \quad \forall i, j \in K \tag{4.3}$$

The quantity of the transferred risk is a function of the difference in $\beta_{x,i}$ between causes, the drift of the trend and the proportion of each cause.

In order to better understand the formula, Figure 4.5 shows the difference of $\beta_{x,i}$ compared to *Drug-related* for male and female for all ages which gives some insights about the direction of risk transfer defined in the equation above. Most of the difference is positive in the young ages, it means that *Drug-related* gained more proportion in younger ages and this observation is more important in male compared to female, which is in line with what has been detected in historical observation.

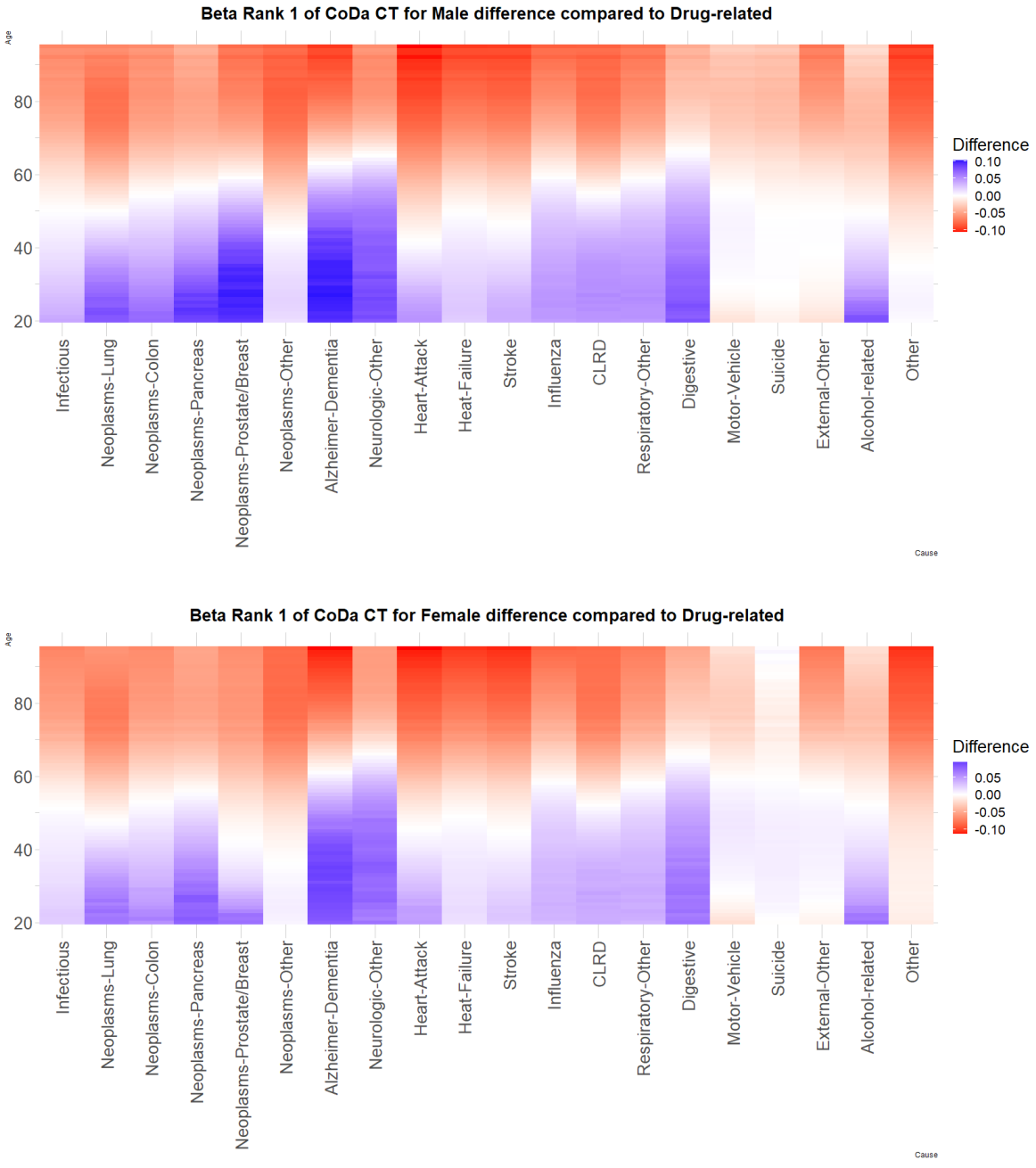


Figure 4.5: CoDa Common Trend $\beta_{x,i}$ difference compared to *Drug-related*

Multi Trend: parameters interpretation and risk transfer concept

While Common Trend assumes a shared trend for each cause, CoDa CoDa Multi Trend considers the evolution dynamics of each cause and thus attributes a trend factor to each cause, and still follows the same constraint. Each cause's clr-transformed matrix of dimension N x T is decomposed separately as shown in Table 4.3, each cause has its own trend parameter $\kappa_{t,i}$ and $\beta_{x,i}$. The interpretation of $\beta_{x,i}$ is thus different in Multi Trend as it represents the age and cause of death trend specific sensitivities.

The formula related to risk transfer described for the CoDa Common Trend Equation 4.3 needs to be further adjusted accordingly for CoDa Multi Trend.

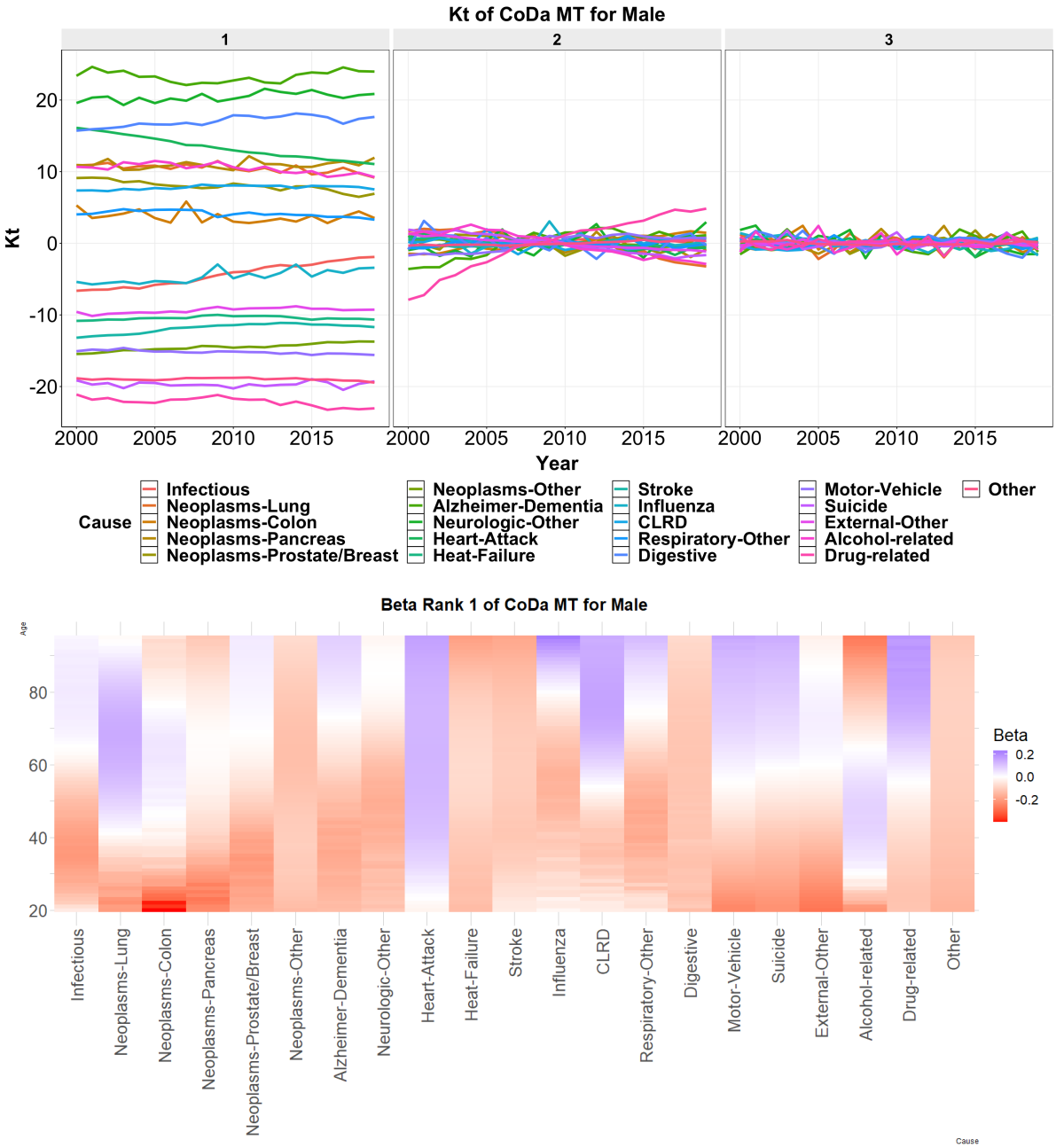


Figure 4.6: CoDa Multi Trend $\kappa_{t,i}$ and $\beta_{x,i}$ for male

It could be seen that according to Figure 4.6 and Figure 4.7 which illustrate the $\kappa_{t,i}$ and $\beta_{x,i}$ of each cause, not all the causes followed the same trend in the past 20 years. Heart attack had a decreasing trend while Influenza and Infectious had a increasing trend. $\beta_{x,i}$ distribution between male and female has the same observation than in the CoDa

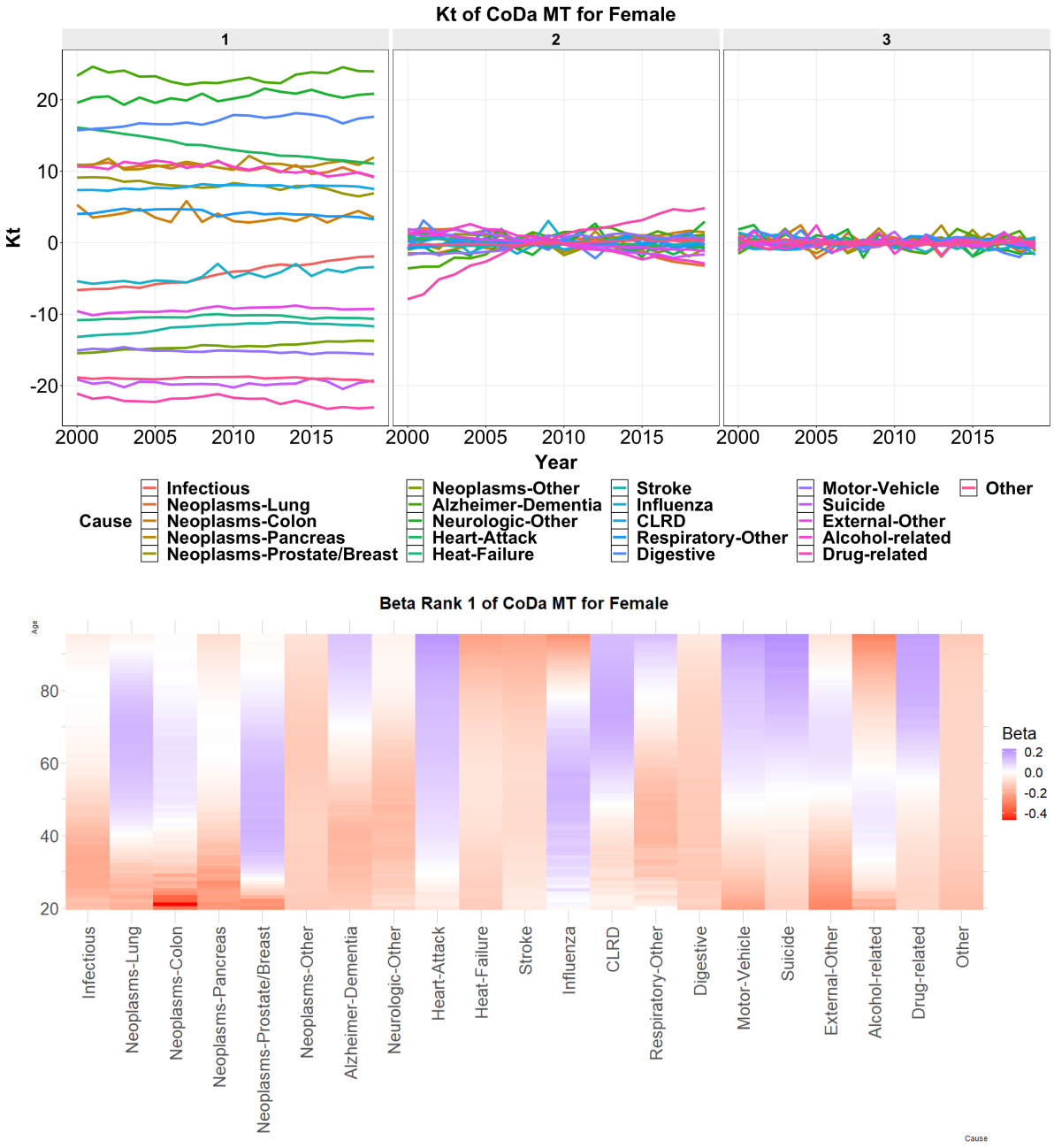


Figure 4.7: CoDa Multi Trend $\kappa_{t,i}$ and $\beta_{x,i}$ for female

Common Trend, most of the causes follow the same distribution, except for *Neoplasm - Prostate/Breast* which has different target age groups.

4.3.2 Model forecast

With the aim of long term forecasting using CoDa models, both CoDa Common Trend and CoDa Multi Trend are set to forecast until 60 years.

In order to remain comparable with classical All-Cause modelling and previous independent cause-specific model, each component of κ_t ($\kappa_{t,i}$) is modelled by an ARIMA (0,1,0) and the forecast is realized by extrapolation on the time index. Along with the predicted κ_t ($\kappa_{t,i}$), multiplied with previously estimated $\beta_{n,i}$, a forecast clr-transformed matrix $Y^{Forecast}$ could be obtained and inverse transformed to cause-specific proportions: $R^K \Rightarrow S^K$.

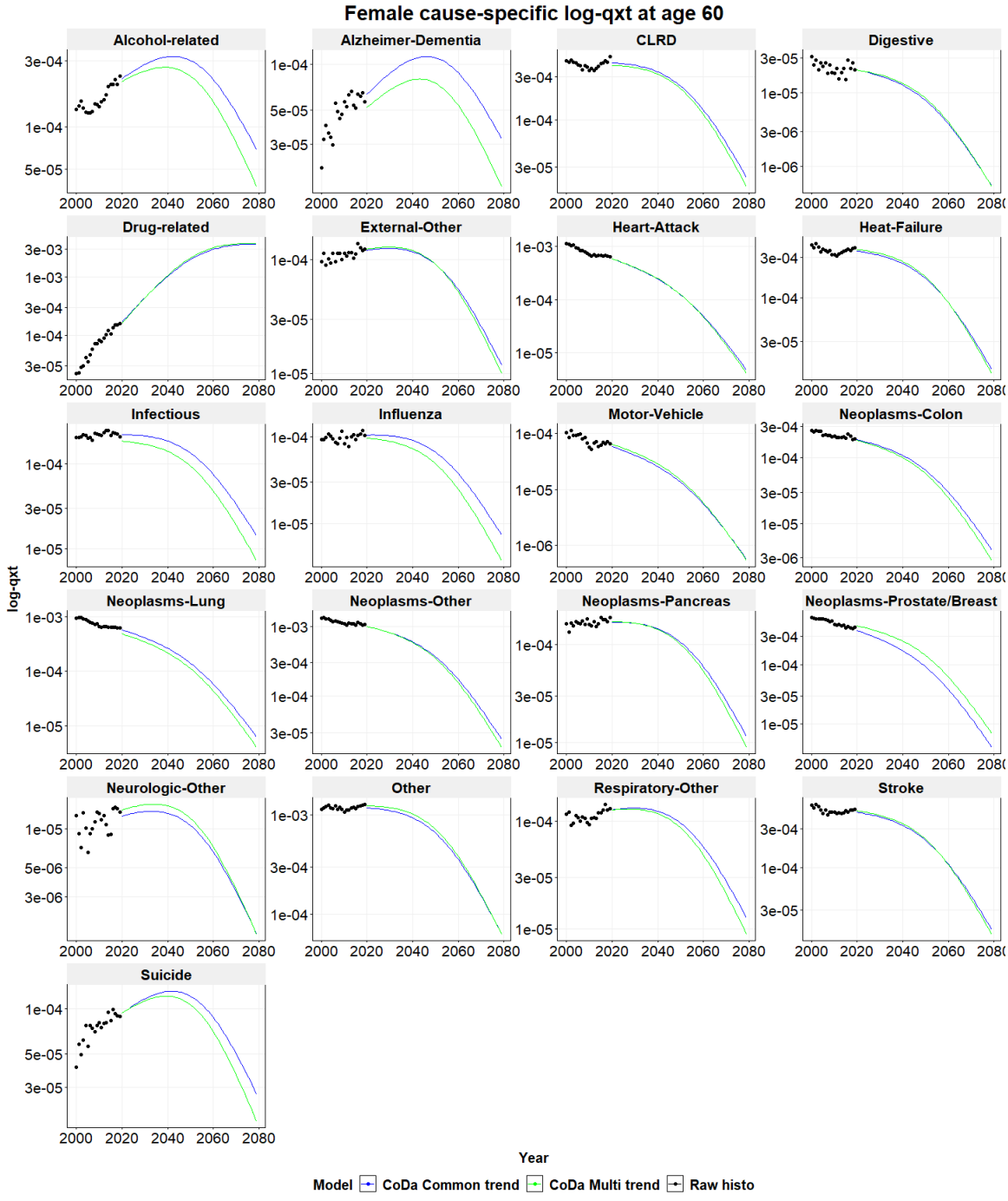


Figure 4.8: Female cause-specific forecast at age 60

Once obtained the forecast cause-specific proportions, cause-specific mortality rates forecast could be calculated by multiplying proportions with the pre-determined All-Cause mortality rates forecast.

Figure 4.8 and Figure 4.9 illustrate cause-specific logarithm mortality rates forecast at age 60 (see Appendix 4.5.1 for more ages), it could be noted that CoDa models generally have a less drastic forecast compared to the independent cause-specific model which extrapolates log-linear historical trend, and shows an opposite trend in certain causes. This is due to the sum constraint, and the fact that the increase of a cause also induces the decrease of others as it can be observed in *Drug-related*. Indeed, all the other causes except *Drug-related* will decrease after 2040.

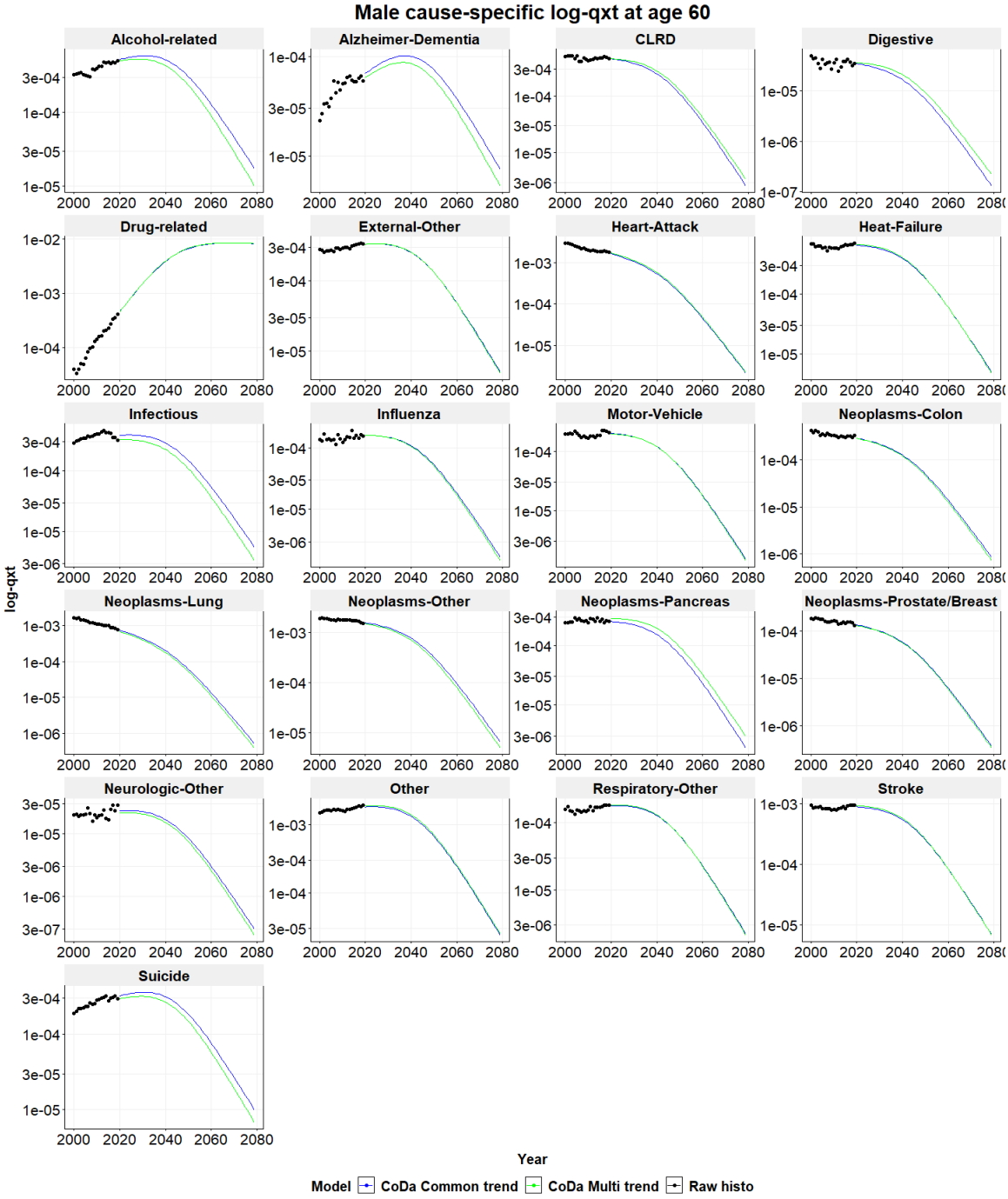


Figure 4.9: Male cause-specific forecast at age 60

4.4 Simulation of scenarios

Central trajectories for each cause have been calculated in the previous section. The next step is to build a prediction interval of aggregate mortality rates, which requires simulating future scenarios.

This section investigates how to simulate future scenarios of aggregate mortality rates within the two CoDa models. The aggregate mortality rates forecast is pre-determined using the classical All-Cause model and under the sum constraint, the sum of cause-specific mortality rates forecast remains the same as with the All-Cause model.

4.4.1 Methodology

Before deriving the aggregate mortality rate dynamics, future deviations of cause specific proportions need to be calculated as they result from the clr transformed values of the compositional Lee-Carter model.

The uncertainties on the future cause-specific proportions are mainly influenced by the trend factor, and the fluctuation of the trend factor could be simulated by adding a yearly deviation e_t based on a run-off view. Since κ_t in both CoDa Common Trend and CoDa Multi Trend are modelled and forecasted by an ARIMA (0,1,0): $\kappa_{t+1} = \kappa_t + \delta + \epsilon_t$, the simulation is based on simulating the residuals ϵ_t which are assumed to follow a normal distribution $\mathcal{N} \sim (0, \sigma^2)$, σ as the standard error of ϵ_t.

Since there exists a unique trend factor κ_t in the CoDa Common Trend, the steps consist of simulating 60 yearly deviations and adding them to the initial forecast κ_t:

$$\kappa_t^{scenario} = \kappa_t^{initial} + e_t$$
$$\kappa_{t+1}^{scenario} = \kappa_t^{initial} + \delta + e_t + e_{t+1}$$

Each simulation contains 60 yearly deviations which cover the whole forecast horizon. New $\kappa_t^{scenario}$ will be multiplied by $\beta_{x,i}$ as well and a new clr-transformed matrix $Y^{scenario}$ is obtained. For the sake of not introducing too much uncertainty from the estimation error, the simulation is only applied to the first rank of κ_t, the rest remain the same as in the initial forecast.

As for the CoDa Multi Trend, the same method from the independent model is used, a correlation matrix between residuals of each $\kappa_{t,i}$ is obtained, Figure 4.10 and Figure 4.11 show the correlation matrix for both genders, from which each residual is assumed to follow a normal distribution and with the hypothesis of a multivariate normal distribution. Each simulation produces 60 yearly deviations for each $\kappa_{t,i}$, followed by $\kappa_{t,i}^{scenario}$ multiplying by $\beta_{x,i}$.

After the 10,000 simulations described above, 10,000 clr-transformed matrices $Y^{scenario}$ are obtained for the CoDa Common Trend and CoDa Multi Trend . In the CoDa modelling approach, these clr-transformed matrices will be returned back to the cause-specific proportion matrix through inverse transformation clr^{-1}.

This thesis proposes a method to adapt the inverse transformation translate the deviations in terms of clr values into deviations in terms of mortality rates. The inverse transformation operator is defined as C(exp(Y)) = $[\frac{exp(Y_i)}{\sum exp(Y_i)}, ..., \frac{exp(Y_j)}{\sum exp(Y_i)}]$

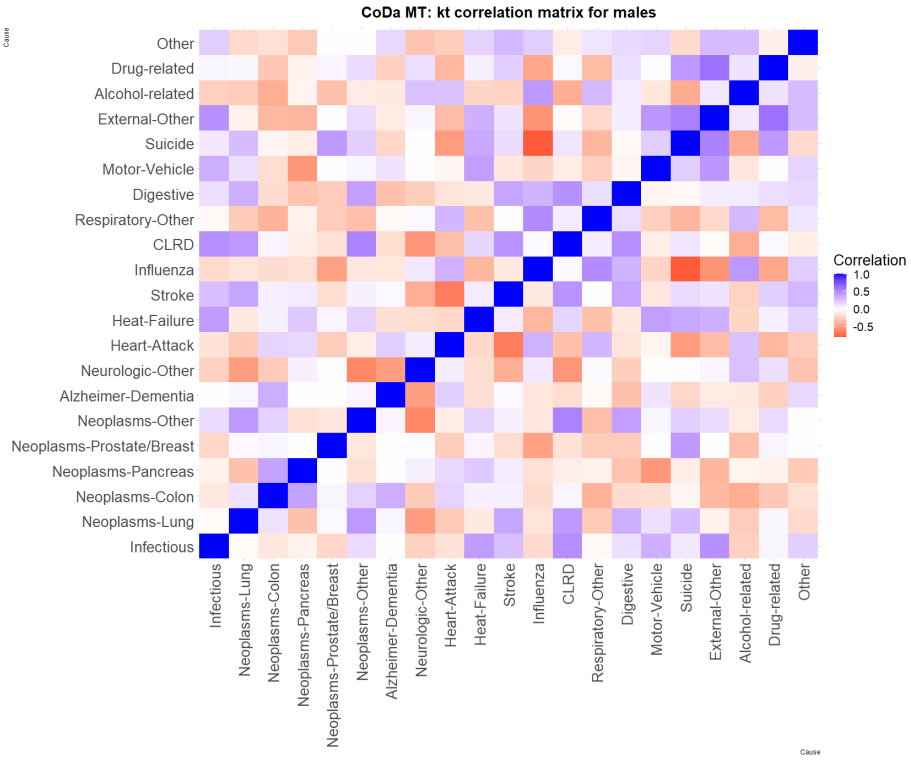


Figure 4.10: CoDa Multi Trend correlation matrix for male

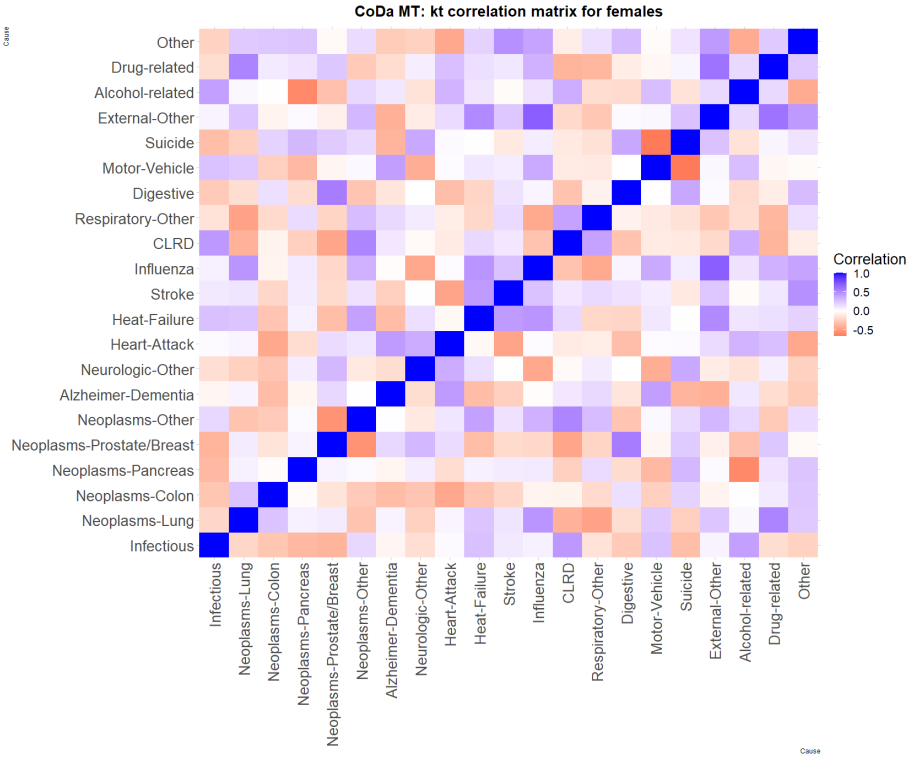
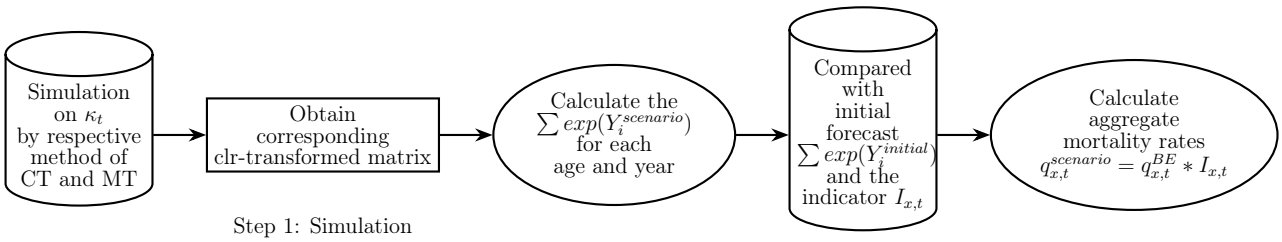


Figure 4.11: CoDa Multi Trend correlation matrix for female



Step 1: Simulation

Step 2: Calculation of $Q_{x,t}^{scenario}$

The denominator $\sum exp(Y_i)$ could be seen as the proxy of aggregate mortality rates, therefore an indicator which is the ratio of this proxy between a given simulation and the central trajectory: $I_{x,t} = \frac{\sum exp(Y_{x,t,i}^{scenario})}{\sum exp(Y_{x,t,i}^{initial})}$. It would represent the evolution of the aggregate mortality rates in percentage for someone aged x and in year t.

4.4.2 Results

After computing the $q_{x,t}^{scenario}$ from the method developed above for each simulation, 10,000 scenarios are obtained for each model, Figure 4.12 and Figure 4.13 illustrate the first 5 scenarios obtained.

It has come to attention that volatility in young and middle-aged groups is greater than in elder age groups, which represents the relative instability of the dependence structure among young and middle-aged groups, elder age groups may encounter more diverse causes of death, which implies a more stable cause dependence structure.

It is noteworthy that female at age 40-50 present more volatilities between CoDa Common Trend and CoDa Multi Trend, and between male and female. The explanation is that since CoDa Multi Trend considers the cause-specific evolution, it results in more uncertainties on the future aggregate trajectories. Furthermore, as observed before, there exist causes which have different target ages between male and female. For instance, *Neoplasm-Prostate/Breast* impacts younger ages for female.

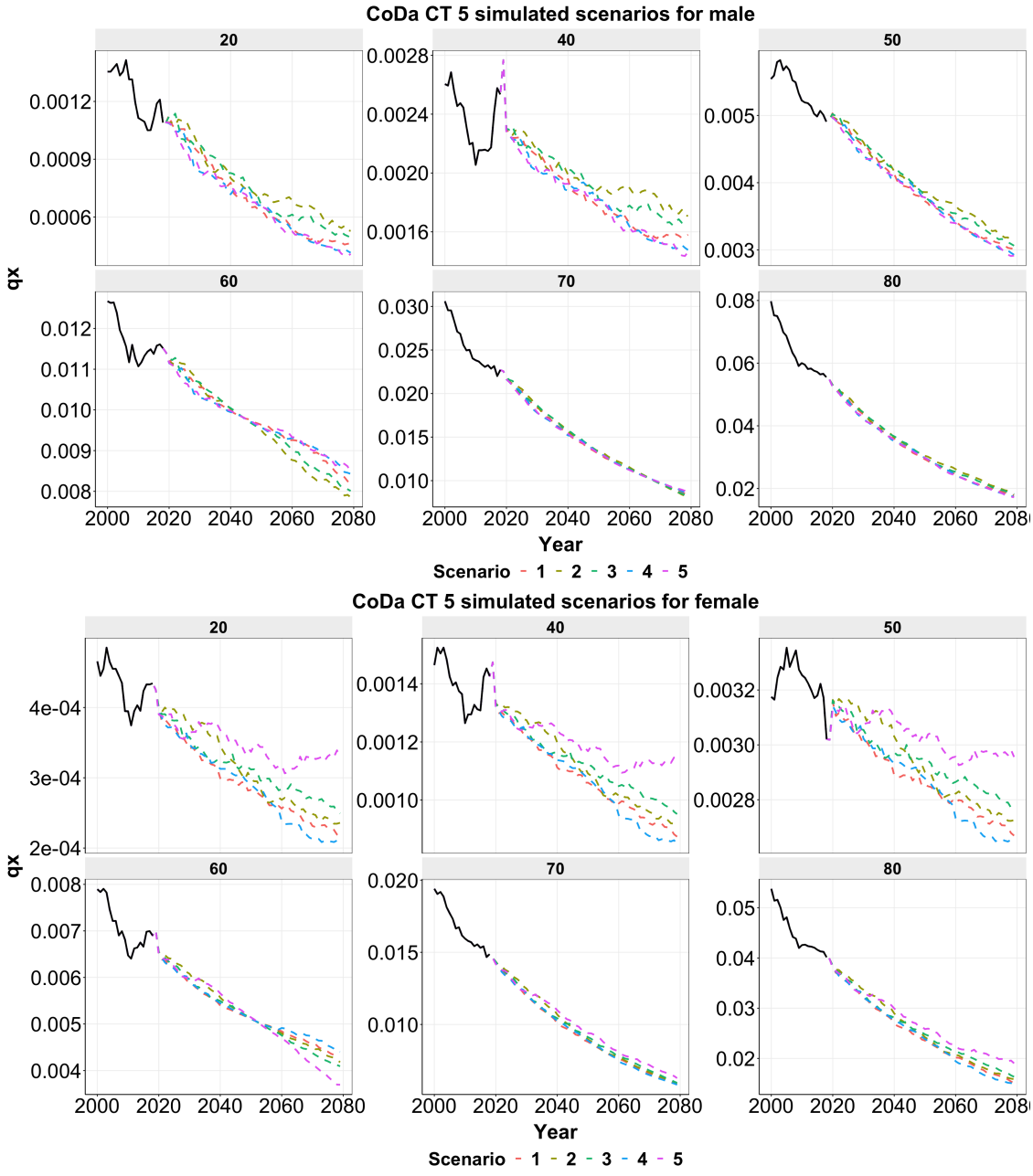


Figure 4.12: CoDa Common Trend 5 scenarios

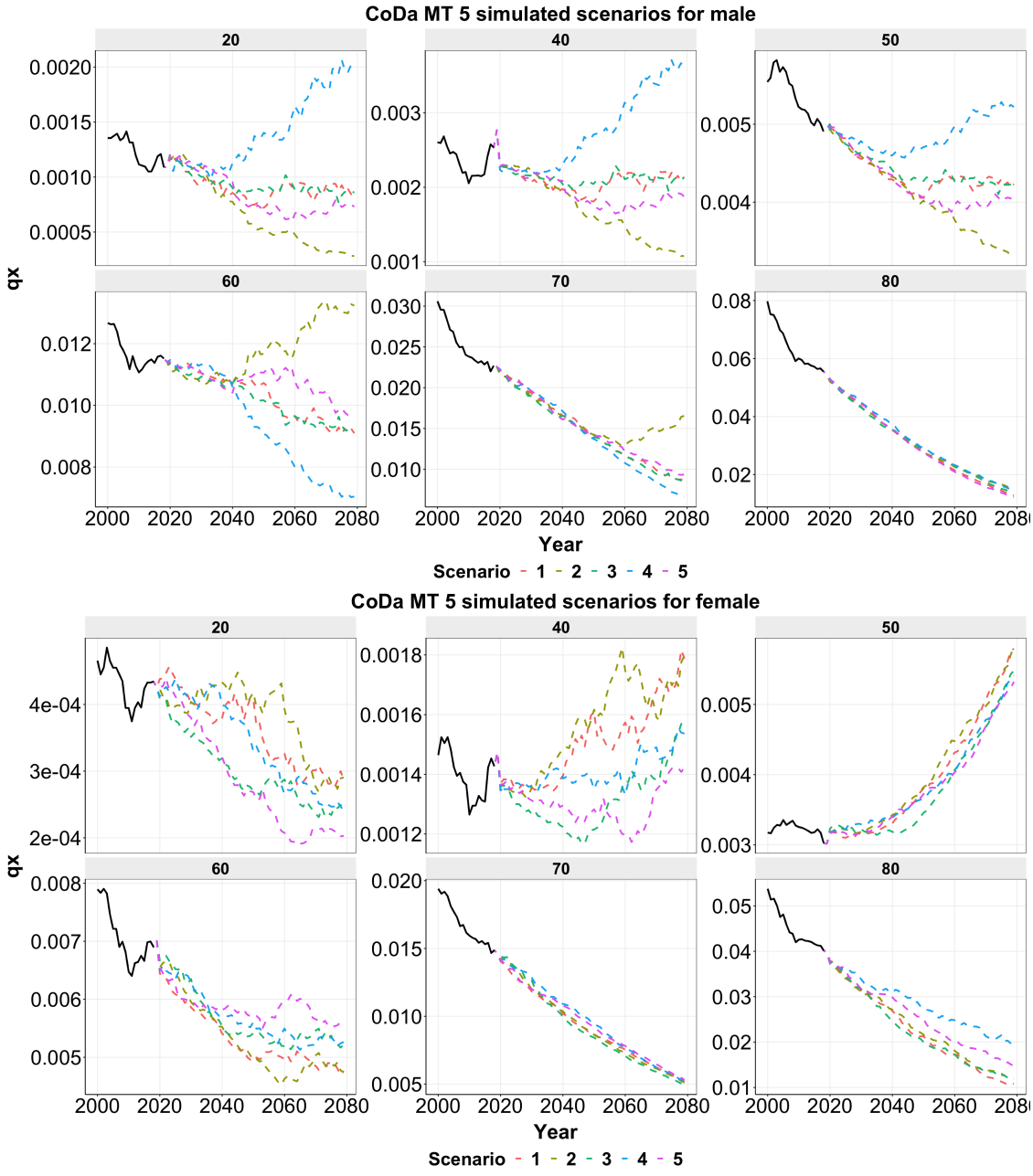


Figure 4.13: CoDa Multi Trend 5 scenarios

4.5 Life expectancy

This section aims to evaluate cohort life expectancy according to the simulated scenarios. Cohort life expectancy is suggested as a mortality sensitivity measure in the internal model from EIOPA [2018]. Based on the initial age range calibrated: 20-95, cohort life tables are constructed by assuming the last year of life at 95-year-old and extending the forecast horizon to 76 years.

In this section, no closing table methods were employed since cause-of-death data quality at extreme ages may not be satisfactory, some methodologies regarding interpolation or CoDa itself could be used to resolve cause-of-death data problem in extreme ages [Piveteau [2021]]. Expected age at death is defined at attained age in 2020 plus corresponding cohort life expectancy, Figure 4.15 and Figure 4.16 demonstrate the expected age at death for a male/female who attained age from 20 to 80 in 2020 which is the first year of projection as the data is from 2000 to 2019. The age range choice is based on the exposure distribution according to the 2017 Individual Life Insurance Mortality Experience Report from Society of Actuaries (SoA). ILEC is a committee of the SOA (Society of Actuaries) that gathers the information necessary for the construction of the VBT (Valuation Basic Table) mortality tables. It could be seen that age 20-80 occupy more major proportions over 90% in total.

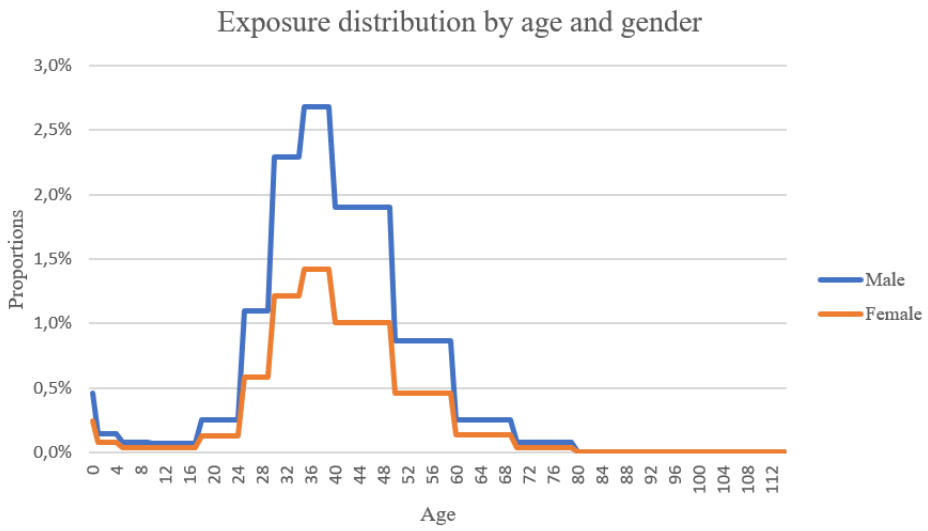


Figure 4.14: Exposure distribution by age and gender

Expected age at death prediction interval at 99% from All-Cause and CoDa models are also shown in the same figures. The cohort life expectancy has been calculated using the same methodology for the three models: All-Cause and CoDa models.

In comparison, the cohort life expectancy of the CoDa Common Trend end up obtaining a narrower interval compared to All-Cause model. The underlying explanation is not straightforward because the models are different as shown below:

$$\begin{aligned} \text{All-Cause} &: ln(\mu_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t} \\ \text{CoDa Common Trend} &: clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i} \kappa_t + \epsilon_{x,t,i} \\ \text{CoDa Multi Trend} &: clr(d_{x,t,i} \ominus \alpha_{x,i}) = \beta_{x,i} \kappa_{t,i} + \epsilon_{x,t,i} \end{aligned}$$

The deviation of $q_{x,t}$ depends on both β_x and the standard error of $\epsilon_{t,i}$, the scale of κ_t and the meaning of β_x are different between the three models. Furthermore, Table 4.6 shows the standard error of residuals ϵ_t in the κ_t based on which yearly deviations e_t are

	All-Cause for female	CoDa Common Trend for female	All-Cause for male	CoDa Common Trend for male
Standard error of κ_t	0.501	0.431	0.507	0.482

Table 4.6: Standard error of κ_t

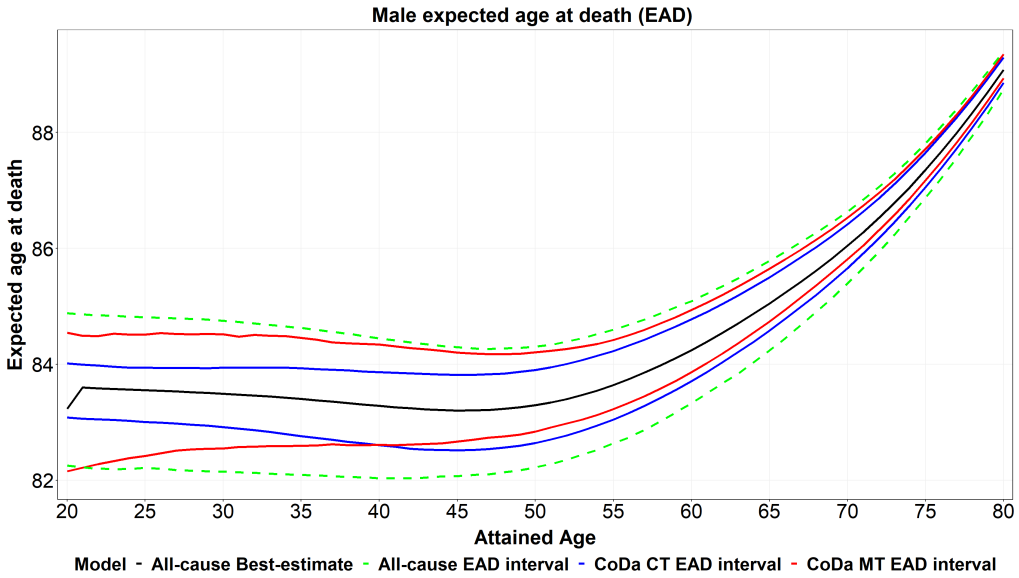


Figure 4.15: Male cohort expected age at death interval

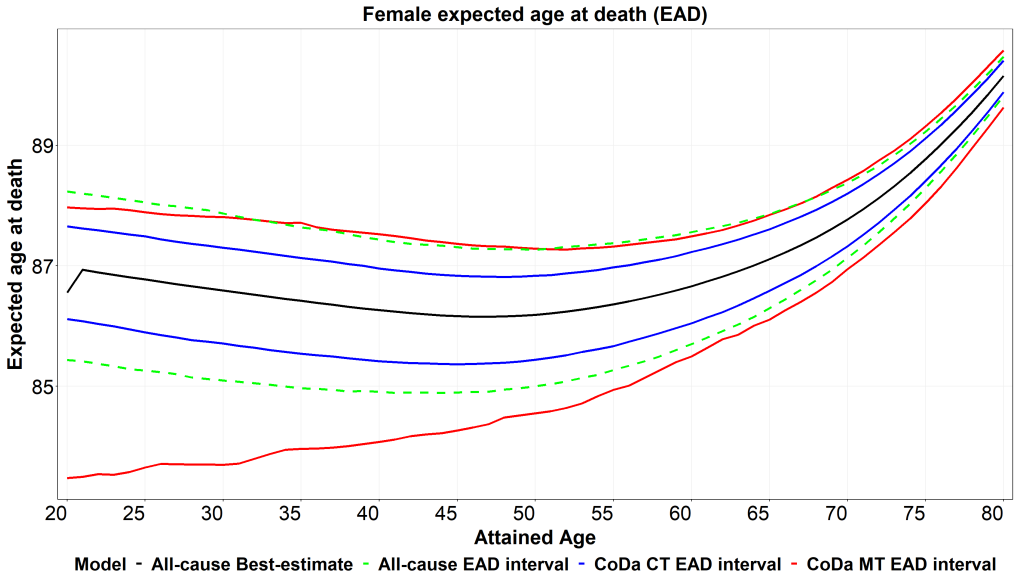


Figure 4.16: Female cohort expected age at death interval

simulated, CoDa Common Trend presented less volatility on κ_t, which could partly explain the reason why CoDa Common Trend has narrower interval compared to All-Cause model.

Between CoDa Common Trend and CoDa Multi Trend, it can be observed that the deviations are more important for the younger ages and lower for the elder ages for CoDa Multi Trend than CoDa Common Trend. Indeed, CoDa Multi Trend also takes into account the different evolution of each cause and consequently creates more uncertainties.

It can also be seen that CoDa Multi Trend presents a larger interval for female, especially at younger ages, which is in line with the observation in the simulated scenarios. Due

to the different age ranges impacted by certain causes such as *Neoplasms-Prostate/Breast*, female display more volatility among young and middle ages.

4.5.1 Scenario analysis

The 0,5% scenario (which represents actually a 99,5% scenario in terms of Solvency II shock) has been further analysed. The interpretability ability of CoDa models has been explored by analyzing the scenario corresponding to the 0.5% level of cohort expectancy at age 20, as it can be insightful to explore each cause's proportion evolution throughout the forecast horizon.

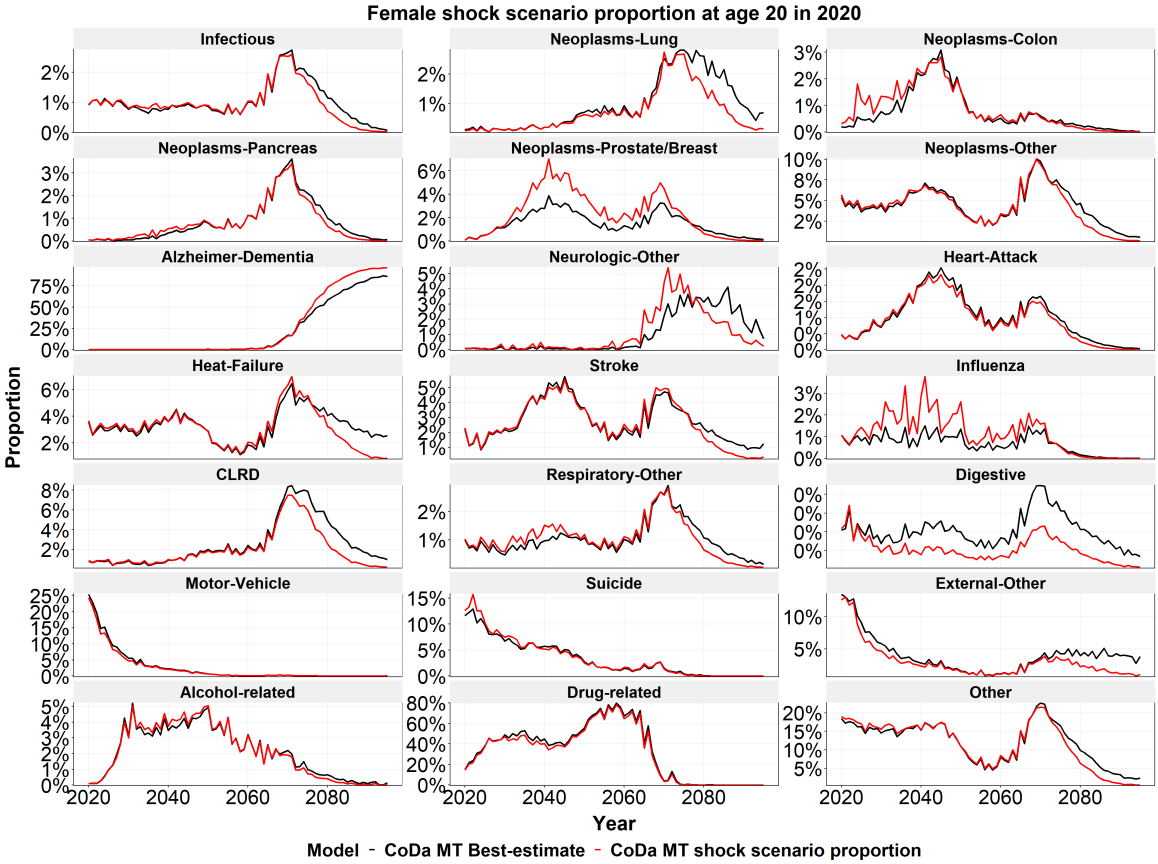


Figure 4.17: Best-estimate and Expected age at death at 0.5% level scenario of female cohort aged 20 in 2020

Figure 4.17 and Figure 4.18 illustrate the cause proportion evolution for the female and male cohort aged 20 in 2020, for the 0,5% level scenario, i.e. the scenario which represents the 0,5% quantile for the expected age at death is identified and analysed. It can be observed that a significant increase of proportion in *Neoplasm-Prostate/Breast* for female before 2070, therefore before the cohort reaches the age of 70. This explains the discrepancy between the life expectancy of the scenario and the central trajectory. Above the age 70 (the year 2070), causes related to neurologic system including *Alzheimer-Dementia* and *Neurologic-Other* are the main source of the decline in female life expectancy.

Also, *Drug-related* and *Alzheimer-Dementia* occupy a major proportion after 40 years for both female and male respectively in young-middle-ages and elder ages, this also shows one of the limitations of CoDa model that despite coherent cause-specific mortality rates forecast, certain cause-specific proportion may be dominant as well in long term forecast horizon under the linear extrapolation method.

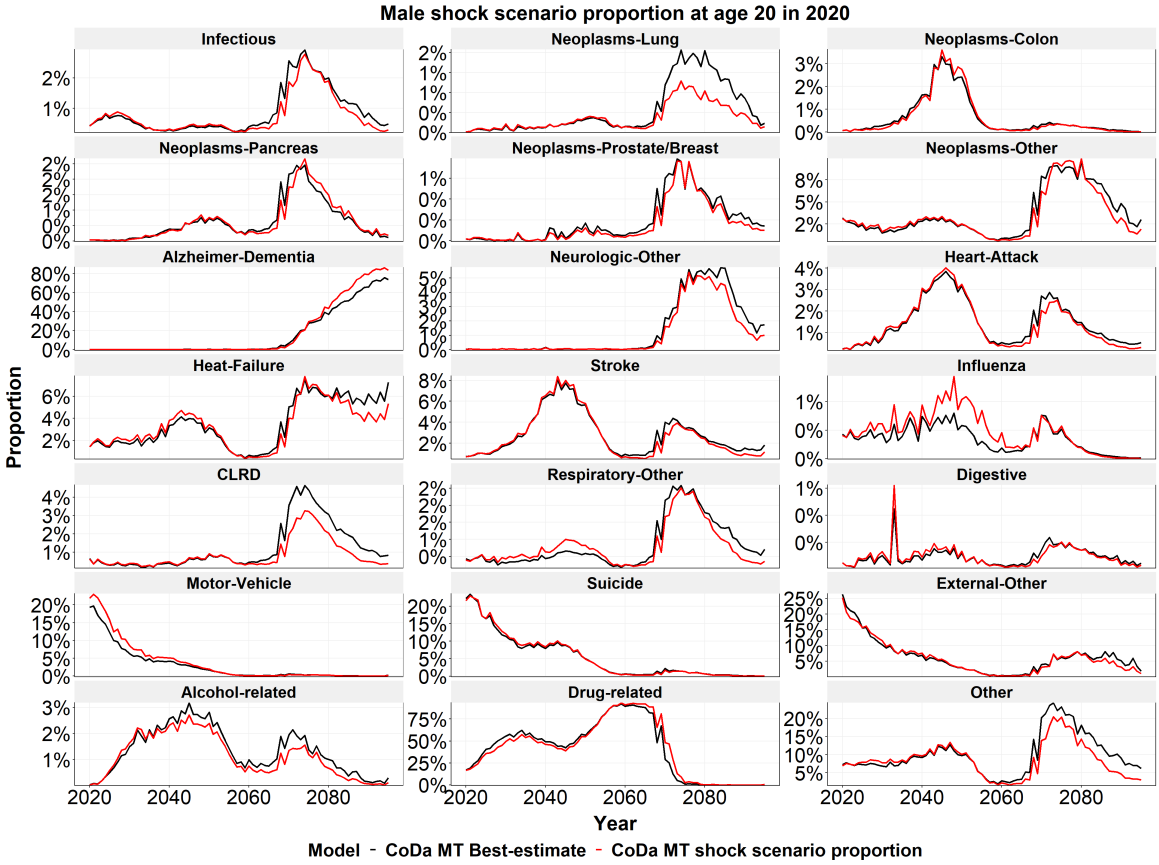


Figure 4.18: Best-estimate and Expected age at death at 0.5% level scenario of male cohort aged 20 in 2020

Conclusion

This thesis explored the cause-of-death in a mortality risk modelling framework with three modelling approaches and compared their respective advantages and limits regarding the interpretability of the observed mortality experience, and suitability compared to the common practice of mortality risk calibration.

The first tentative is to model each cause-of-death independently, which proves satisfactory interpretability on the historical trend of each cause. Forecast by a linear extrapolation demonstrates incoherence on the aggregate mortality rates' evolution, less visible in the short term and more evident for long-term projection, as required in long-term mortality risk modelling

As an alternative approach, this thesis then investigated Compositional Data analysis (CoDa) application on cause-of-death modelling, which provides a more coherent cause-specific forecast by imposing an aggregate level constraint, two variants of the CoDa model were tested which all illustrate the capacity to output more coherent result.

In order to simulate future scenarios and build a prediction interval of aggregate mortality rates, this thesis then proposed a method to impact aggregate level mortality rates by analyzing cause-specific evolution uncertainties. CoDa models have shown more coherent long-term cause-specific mortality rates forecast with respect to aggregate mortality rates, their explanatory ability shed light on the contribution of each cause in the mortality risk and it helps to understand the risk transfer mechanism within ages and causes.

However, certain causes such as *Drug-related* and *Alzheimer-Dementia* could still be dominant in proportions in the long-term forecast. According to CDC, from 2018 to 2019, drug overdose deaths increased by nearly 5%, quadrupling since 1991. Over 70% of 70,630 deaths in 2019 were opioid-related. The same observation is found in the models employed in the thesis.

Recent public policies have started to remedy drug-related overdose and companies selling opioids have been sued in the USA. This impact should be considered in the modelling approach.

Hence, this thesis could be further deepened, by applying expert judgement on each cause-of-death to limit their evolution on a more rational scale. The impact on mortality risk could consequently be defined as the result of the application of expert judgement. Based on the CoDa models, one could obtain a coherent cause-specific forecast, in which one could apply future mortality improvement correction from expert judgement. Another solution could be to set up an upper boundary on cause-specific mortality rates, the aggregate mortality rates will consequently change. The use of expert judgement shall be objectively justified and is out of the scope of this thesis, hence the use of expert judgement has not been discussed.

Some fundamental issues related to cause-of-death modelled remain to be solved, such as data quality following the change of classification standard, deaths number of each cause may not be stable due to the change of classification.

In conclusion, the direct application of cause-of-death modelling is deemed to be

premature without further adjustment and additional analysis, for the mortality risk assessment. However, it can provide valuable insight on mortality trends and their evolution, detect the main drivers of aggregated mortality risk and analyse extreme scenarios.

Bibliography

WHO ICD-11 Tooling, howpublished = https://www.who.int/standards/classifications/classification-of-diseases/cause-of-death, note = Accessed: 2022-08-30.

Pwc Solvency II Life Insurers' Capital Model Survey. 2019.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

D. H. Alai, S. Arnold, M. Bajekal, and A. M. Villegas. Mind the gap: a study of cause-specific mortality by socioeconomic circumstances. *North American Actuarial Journal*, 22(2):161–181, 2018.

M.-P. Bergeron-Boucher and S. Kjærgaard. Mortality forecasts by age and cause of death: How to forecast both dimensions? SocArXiv d7hbp, Center for Open Science, June 2022. URL `https://ideas.repec.org/p/osf/socarx/d7hbp.html`.

M.-P. Bergeron-Boucher, M. Ebeling, and V. Canudas-Romo. Decomposing changes in life expectancy: Compression versus shifting mortality. *Demographic Research*, 33:391–424, 2015.

M.-P. Bergeron-Boucher, V. Canudas-Romo, J. Oeppen, and J. W. Vaupel. Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, 37: 527–566, 2017.

A. Boumezoued, J.-B. Coulomb, A. Klein, D. Louvet, and E. Titon. Modeling and forecasting cause-of-death mortality. *Society of actuaries (SOA)*, 2019. URL `https://www.soa.org/resources/research-reports/2019/cod-mortality-forecasting/`.

N. Brouhns, M. Denuit, and J. K. Vermunt. A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and economics*, 31 (3):373–393, 2002.

R. L. Brown and J. McDaid. Factors affecting retirement mortality. *North American Actuarial Journal*, 7(2):24–43, 2003.

A. J. Cairns, D. Blake, and K. Dowd. A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4): 687–718, 2006.

CDC. Understanding the epidemic. URL `https://www.cdc.gov/drugoverdose/epidemic/index.html`.

P. Cox. Mathematical models for the growth of human populations. *Journal of the Institute of Actuaries*, 99(3):307–308, 1973.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

EIOPA. EIOPA's second set of advice to the European Commission on specific items in the Solvency II Delegated Regulation. 2018.

K. K. Elizabeth Arias, Betzaida Tejada-Vera and F. B. Ahmad. Provisional life expectancy estimates for 2021. *The National Center for Health Statistics*, 2022. URL extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cdc.gov/nchs/data/vsrr/vsrr023.pdf.

E. European commission. Solvency II. URL https://www.eiopa.europa.eu/browse/solvency-2_en#2020SolvencyIIreview.

N. Huynh and M. Ludkovski. Joint models for cause-of-death mortality in multiple populations. *arXiv preprint arXiv:2111.06631*, 2021.

Institute and Faculty of Actuaries. Solvency II - Life Insurance . 2016.

M. W. Jayawardana and G. Sofronov. Multiple break-points detection in array cgh data via the cross-entropy method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11 2014. doi: 10.1109/TCBB.2014.2361639.

S. Kjærgaard, Y. E. Ergemen, M. Kallestrup-Lamb, J. Oeppen, and R. Lindahl-Jacobsen. Forecasting Causes of Death using Compositional Data Analysis: the Case of Cancer Deaths. CREATES Research Papers 2019-07, Department of Economics and Business Economics, Aarhus University, May 2019. URL https://ideas.repec.org/p/aah/create/2019-07.html.

R. D. Lee and L. R. Carter. Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992. ISSN 01621459. URL http://www.jstor.org/stable/2290201.

H. Li and Y. Lu. Modeling cause-of-death mortality using hierarchical archimedean copula. *Scandinavian Actuarial Journal*, 2019(3):247–272, 2019. doi: 10.1080/03461238.2018.1546224. URL https://doi.org/10.1080/03461238.2018.1546224.

M. Ludkovski, J. Risk, and H. Zail. Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin: The Journal of the IAA*, 48(3):1307–1347, 2018.

E. National Academies of Sciences and Medicine. *The Growing Gap in Life Expectancy by Income: Implications for Federal Programs and Policy Responses*. The National Academies Press, Washington, DC, 2015. ISBN 978-0-309-31707-8. doi: 10.17226/19015.

E. National Academies of Sciences, Medicine, et al. *High and rising mortality rates among working-age adults*. 2021.

J. Oeppen et al. Coherent forecasting of multiple-decrement life tables: a test using japanese cause of death data. 2008.

E. Parliament and of the Council. Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II) . 2009.

S. Piveteau. *Modélisation de la mortalité par cause de décès*. Theses, Université de Lyon, July 2021. URL `https://tel.archives-ouvertes.fr/tel-03499914`.

S. Piveteau and J. Tomas. Mortality forecasting by cause of death and basis risk modelling with compositional data. Presented at the Longevity 14 Conference, Amsterdam,, 2018. URL `extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.bayes.city.ac.uk/__data/assets/pdf_file/0019/437140/TOMAS-Julian.pdf`.

S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

J. R. Wilmoth. Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies*, 5(4):293–319, 1995.

Appendix A Cause-of-death mapping list

Causes	Code
Infectious	001 = Salmonella infections (A01-A02)
Infectious	002 = Shigellosis and amebiasis (A03, A06)
Infectious	003 = Certain other intestinal infections (A04, A07-A09)
Infectious	005 = Respiratory tuberculosis (A16)
Infectious	006 = Other tuberculosis (A17-A19)
Infectious	007 = Whooping cough (A37)
Infectious	008 = Scarlet fever and erysipelas (A38, A46)
Infectious	009 = Meningococcal infection (A39)
Infectious	010 = Septicemia (A40-A41)
Infectious	011 = Syphilis (A50-A53)
Infectious	012 = Acute poliomyelitis (A80)
Infectious	013 = Arthropod-borne viral encephalitis (A83-A84, A85.2)
Infectious	014 = Measles (B05)
Infectious	015 = Viral hepatitis (B15-B19)
Infectious	016 = Human immunodeficiency virus (HIV) disease (B20-B24)
Infectious	017 = Malaria (B50-B54)
Infectious	018 = Other and unspecified infectious and parasitic diseases and their sequelae (A00, A05, A20-A36, A42-A44, A48-A49, A54-A79, A81-A82, A85.0, A85.1, A85.8, A86, B04, B06-B09, B25-B49, B55-B99)
Neoplasms-Lung	027 = Malignant neoplasms of trachea, bronchus and lung (C33-C34)
Neoplasms-Colon	023 = Malignant neoplasms of colon, rectum and anus (C18-C21)
Neoplasms-Pancreas	025 = Malignant neoplasm of pancreas (C25)
Neoplasms-Prostate/Breast	029 = Malignant neoplasm of breast (C50)
Neoplasms-Prostate/Breast	033 = Malignant neoplasm of prostate (C61)
Neoplasms-Other	020 = Malignant neoplasm of lip, oral cavity and pharynx (C00-C14)
Neoplasms-Other	021 = Malignant neoplasm of esophagus (C15)
Neoplasms-Other	022 = Malignant neoplasm of stomach (C16)
Neoplasms-Other	024 = Malignant neoplasms of liver and intrahepatic bile ducts (C22)
Neoplasms-Other	026 = Malignant neoplasm of larynx (C32)
Neoplasms-Other	028 = Malignant melanoma of skin (C43)
Neoplasms-Other	030 = Malignant neoplasm of cervix uteri (C53)
Neoplasms-Other	031 = Malignant neoplasms of corpus uteri and uterus, part unspecified (C54-C55)
Neoplasms-Other	032 = Malignant neoplasm of ovary (C56)
Neoplasms-Other	034 = Malignant neoplasms of kidney and renal pelvis (C64-C65)
Neoplasms-Other	035 = Malignant neoplasm of bladder (C67)
Neoplasms-Other	036 = Malignant neoplasms of meninges, brain and other parts of central nervous system (C70-C72)
Neoplasms-Other	038 = Hodgkin's disease (C81)
Neoplasms-Other	039 = Non Hodgkin's lymphoma (C82-C85)
Neoplasms-Other	040 = Leukemia (C91-C95)
Neoplasms-Other	041 = Multiple myeloma and immunoproliferative neoplasms (C88, C90)
Neoplasms-Other	042 = Other and unspecified malignant neoplasms of lymphoid, hematopoietic and related tissue (C96)
Neoplasms-Other	043 = All other and unspecified malignant neoplasms (C17, C23-C24, C26-C31, C37-C41, C44-C49, C51-C52, C57-C60, C62-C63, C66, C68-C69, C73-C80, C97)
Alzheimer-Dementia	052 = Alzheimer's disease and dementia (G30, F01, F03)*
Neurologic-Other	050 = Meningitis (G00, G03)
Neurologic-Other	051 = Parkinson's disease (G20-G21)
Heart-Attack	059 = Acute myocardial infarction (I21-122)
Heart-Attack	060 = Other acute ischemic heart diseases (I24)
Heart-Attack	062 = Atherosclerotic cardiovascular disease, so described (I25.0)
Heart-Attack	063 = All other forms of chronic ischemic heart disease (I20, I25.1-I25.9)
Heart-Failure	055 = Acute rheumatic fever and chronic rheumatic diseases (I00-I09)
Heart-Failure	065 = Acute and subacute endocarditis (I33)
Heart-Failure	066 = Diseases of pericardium and acute myocarditis (I30-131, I40)
Heart-Failure	067 = Heart failure (I50)
Heart-Failure	068 = All other forms of heart disease (I26-128, I34-138, I42-149, I51)
Stroke	056 = Hypertensive heart disease (I11)
Stroke	057 = Hypertensive heart and renal disease (I13)
Stroke	069 = Essential hypertension and hypertensive renal disease (I10, I12, I15)
Stroke	070 = Cerebrovascular diseases (I60-169)
Stroke	071 = Atherosclerosis (I70)
Stroke	073 = Aortic aneurysm and dissection (I71)
Stroke	074 = Other diseases of arteries, arterioles and capillaries (I72-178)
Stroke	075 = Other disorders of circulatory system (I80-I99)

Influenza	077 = Influenza (J09-J11)
Influenza	078 = Pneumonia (J12-J18)
CLRD	083 = Bronchitis, chronic and unspecified (J40-J42)
CLRD	084 = Emphysema (J43)
CLRD	085 = Asthma (J45-J46)
CLRD	086 = Other chronic lower respiratory diseases (J44, J47)
Respiratory-Other	080 = Acute bronchitis and bronchiolitis (J20-J21)
Respiratory-Other	081 = Other and unspecified acute lower respiratory infections (J22, U04)
Respiratory-Other	087 = Pneumoconioses and chemical effects (J60-J66, J68)
Respiratory-Other	088 = Pneumonitis due to solids and liquids (J69)
Respiratory-Other	089 = Other diseases of respiratory system (J00-J06, J30-J39, J67, J70-J98)
Digestive	090 = Peptic ulcer (K25-K28)
Digestive	091 = Diseases of appendix (K35-K38)
Digestive	092 = Hernia (K40-K46)
Digestive	096 = Cholelithiasis and other disorders of gallbladder (K80-K82)
Motor-Vehicle	114 = Motor vehicle accidents (V02-V04, V09.0, V09.2, V12-V14, V19.0-V19.2, V19.4-V19.6, V20-V79, V80.3-V80.5, V81.0-V81.1, V82.0-V82.1, V83-V86, V87.0-V87.8, V88.0-V88.8, V89.0, V89.2)
Suicide	125 = Intentional self-harm (suicide) by discharge of firearms (X72-X74)
Suicide	126 = Intentional self-harm (suicide) by other and unspecified means and their sequelae (*U03, X60-X71, X75-X84, Y87.0)
External-Other	112 = Accidents (unintentional injuries) (V01-X59, Y85-Y86)
External-Other	115 = Other land transport accidents (V01, V05-V06, V09.1, V09.3-V09.9, V10-V11, V15-V18, V19.3, V19.8-V19.9, V80.0-V80.2, V80.6-V80.9, V81.2-V81.9, V82.2-V82.9, V87.9, V88.9, V89.1, V89.3, V89.9)
External-Other	116 = Water, air and space, and other and unspecified transport accidents and their sequelae (V90-V99, Y85)
External-Other	118 = Falls (W00-W19)
External-Other	119 = Accidental discharge of firearms (W32-W34)
External-Other	120 = Accidental drowning and submersion (W65-W74)
External-Other	121 = Accidental exposure to smoke, fire and flames (X00-X09)
External-Other	123 = Other and unspecified non-transport accidents and their sequelae (W20-W31, W35-W64, W75-W99, X10-X39, X50-X59, Y86)
External-Other	128 = Assault (homicide) by discharge of firearms (*U01.4, X93-X95)
External-Other	129 = Assault (homicide) by other and unspecified means and their sequelae (*U01.0-*U01.3,*U01.5-*U01.9,*U02, X85-X92, X96-Y09, Y87.1)
External-Other	130 = Legal intervention (Y35, Y89.0)
External-Other	132 = Discharge of firearms, undetermined intent (Y22-Y24)

External-Other	133 = Other and unspecified events of undetermined intent and their sequelae (Y34, Y87.2, Y89.9)
External-Other	134 = Operations of war and their sequelae (Y36, Y89.1)
Alcohol-related	094 = Alcoholic liver disease (K70)
Alcohol-related	095 = Other chronic liver disease and cirrhosis (K73-K74)
Drug-related	122 = Accidental poisoning and exposure to noxious substances (X40-X49)
Other	044 = In situ neoplasms, benign neoplasm and neoplasms of uncertain or unknown behavior (D00-D48)
Other	045 = Anemias (D50-D64)
Other	046 = Diabetes mellitus (E10-E14)
Other	048 = Malnutrition (E40-E46)
Other	049 = Other nutritional deficiencies (E50-E64)
Other	098 = Acute and rapidly progressive nephritic and nephrotic syndrome (N00-N01, N04)
Other	099 = Chronic glomerulonephritis, nephritis and nephropathy not specified as acute or chronic, and renal sclerosis unspecified (N02-N03, N05-N07, N26)
Other	100 = Renal failure (N17-N19)
Other	101 = Other disorders of kidney (N25, N27)
Other	102 = Infections of kidney (NI0-N12, N13.6, N15.1)
Other	103 = Hyperplasia of prostate (N40)
Other	104 = Inflammatory diseases of female pelvic organs (N70-N76)
Other	106 = Pregnancy with abortive outcome (O00-O07)
Other	107 = Other complications of pregnancy, childbirth and the puerperium (O10-O99)
Other	108 = Certain conditions originating in the perinatal period (P00-P96)
Other	109 = Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99)
Other	110 = Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R00-R99)
Other	111 = All other diseases (Residual)
Other	135 = Complications of medical and surgical care (Y40-Y84, Y88)
Other	136 = Enterocolitis due to Clostridium difficile (A04.7)

Appendix B Independent cause-specific model parameters

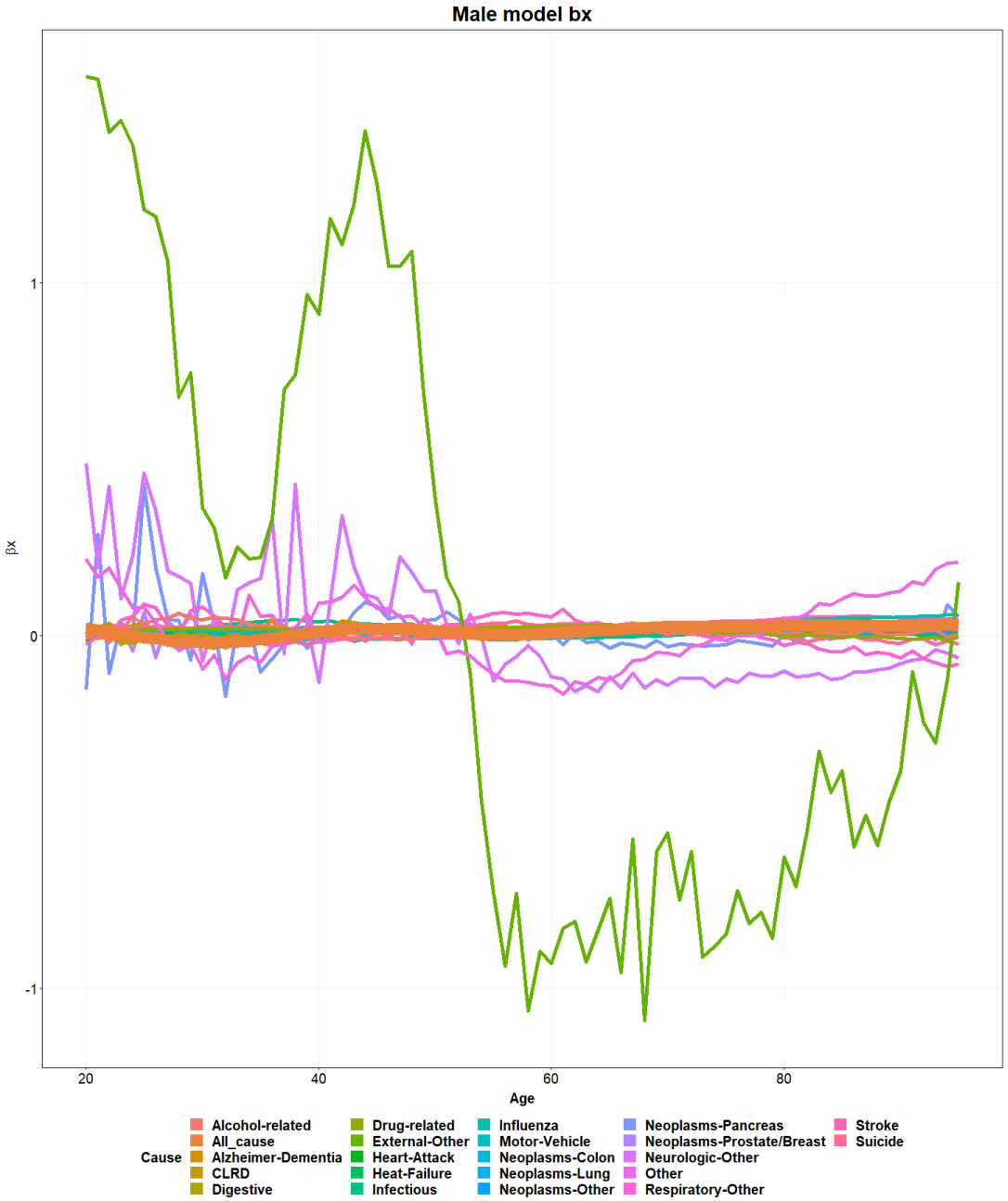


Figure 19: Independent cause-specific model male $\beta_{x,i}$

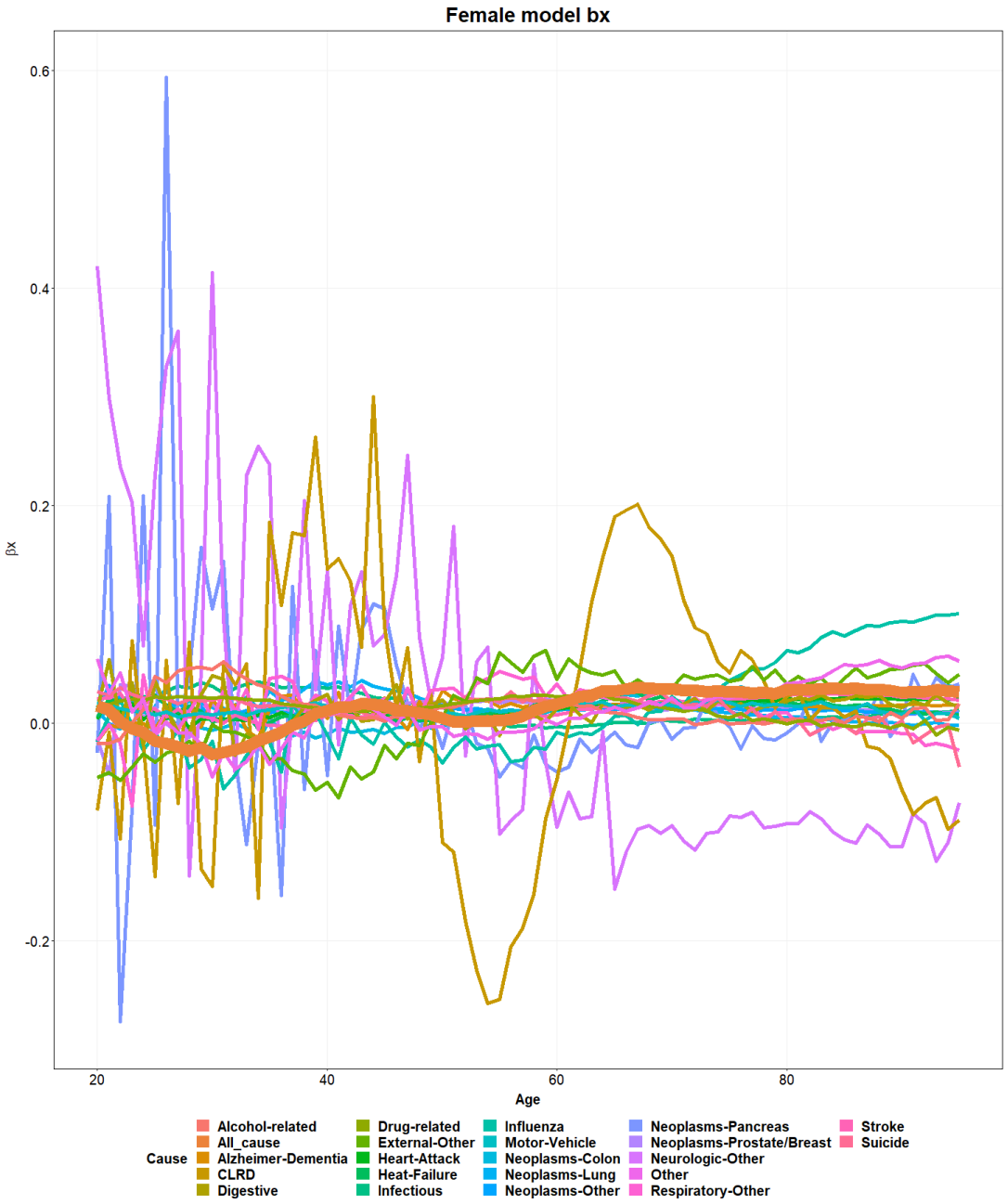


Figure 20: Independent cause-specific model female $\beta_{x,i}$

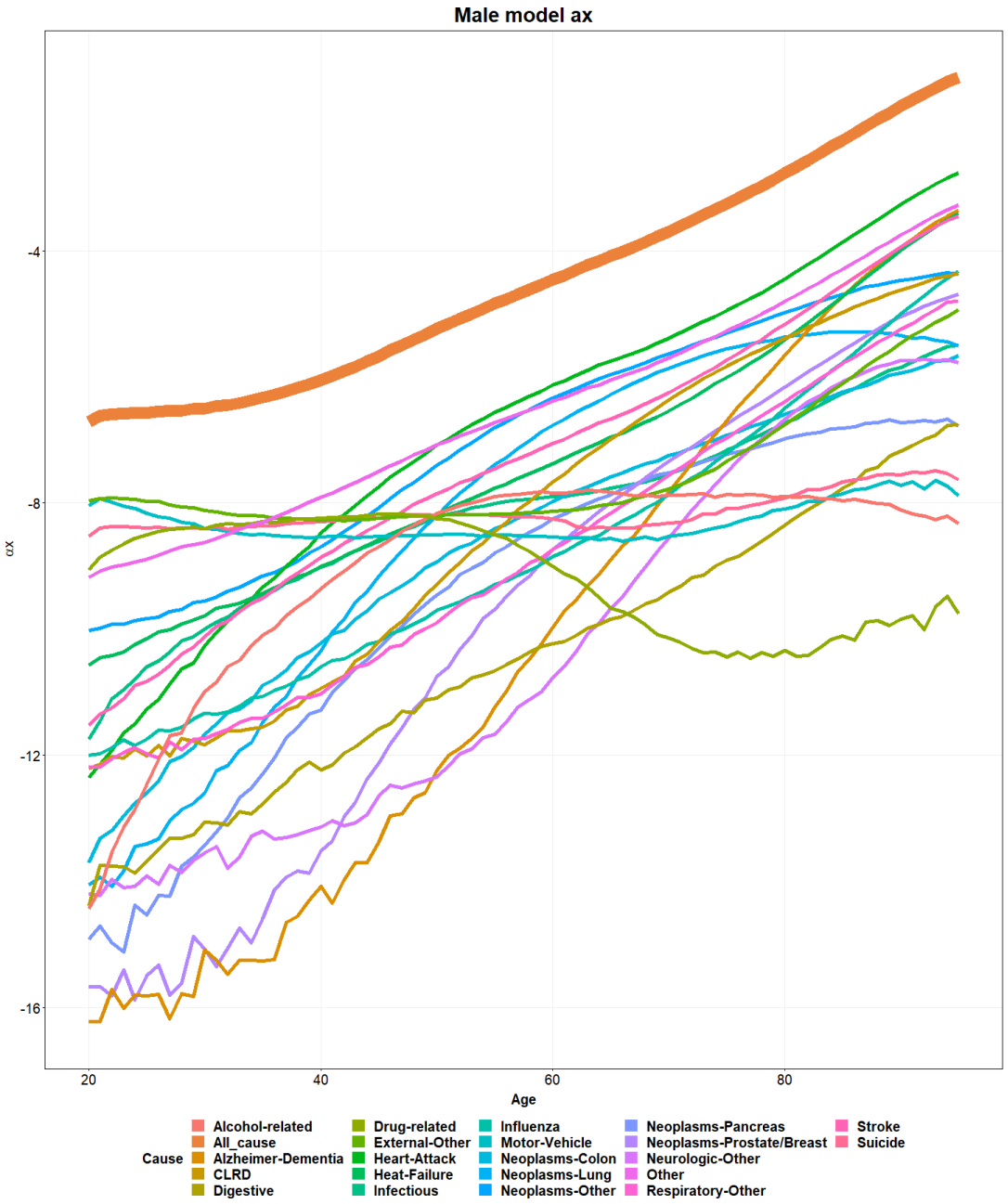


Figure 21: Independent cause-specific model male $\alpha_{x,i}$

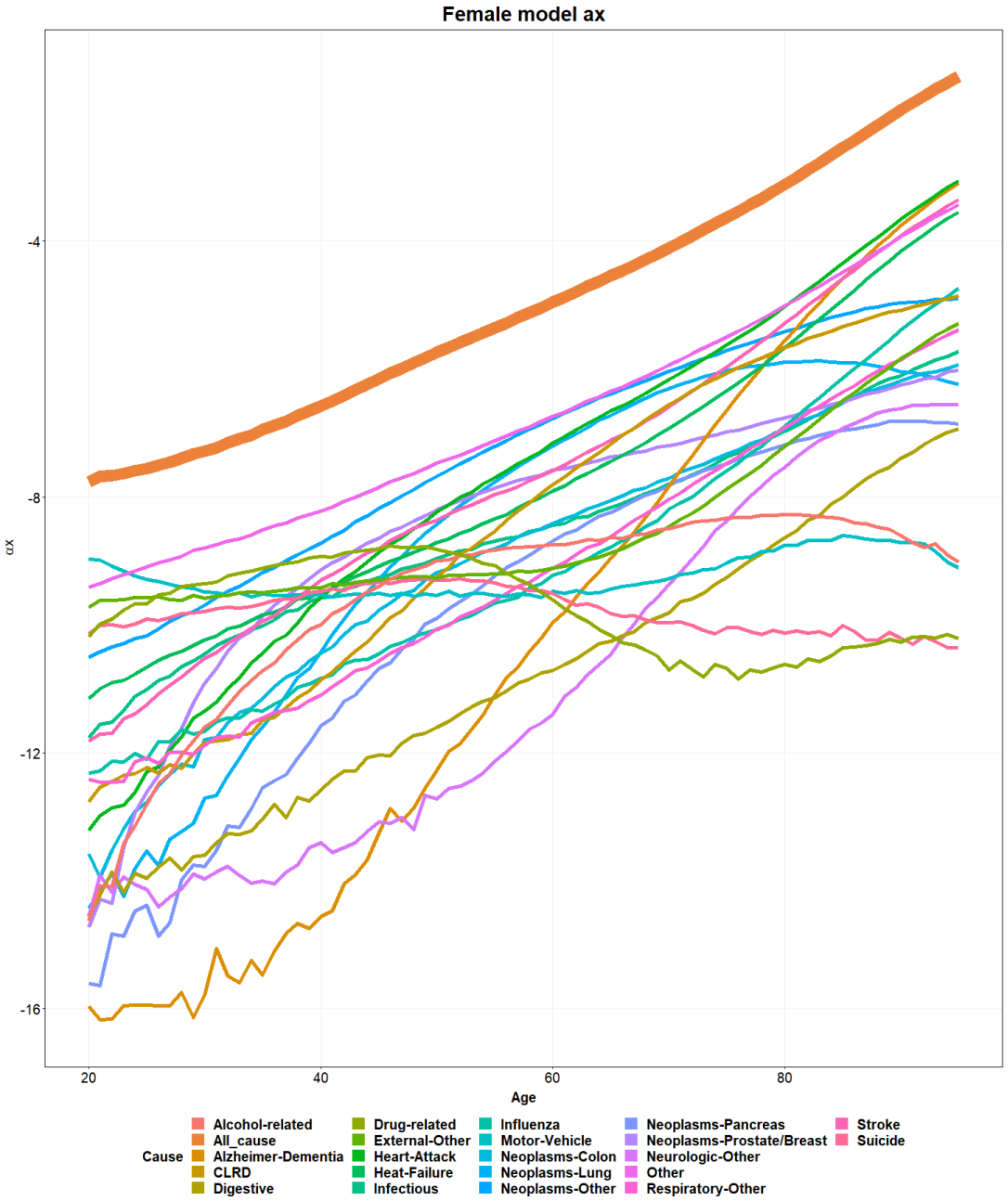


Figure 22: Independent cause-specific model female $\alpha_{x,i}$

According to the parameters $\alpha_{x,i}$, it could be summarized that most of causes follow the same age pattern as in All-cause model except that *Drug-related* demonstrates an inverse major impacted age range. $\beta_{x,i}$ of causes all present more volatility than All-cause model in particular *Drug-related*,especially in younger ages. Apart from the statistical uncertainty of parameters, the low deaths number of cause in young ages is also subject to this effect.

Appendix C CoDa model forecast

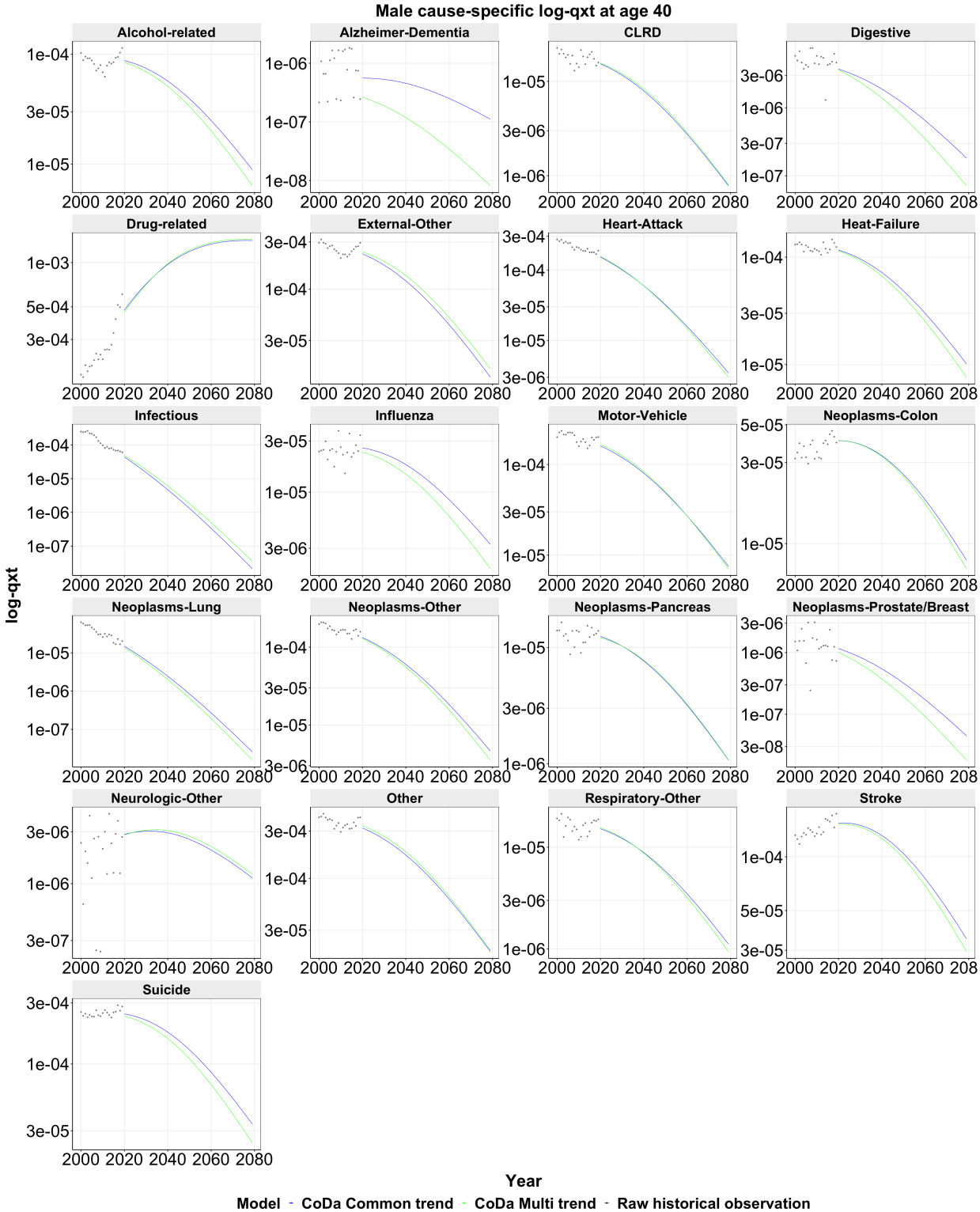


Figure 23: Male cause-specific forecast at age 40

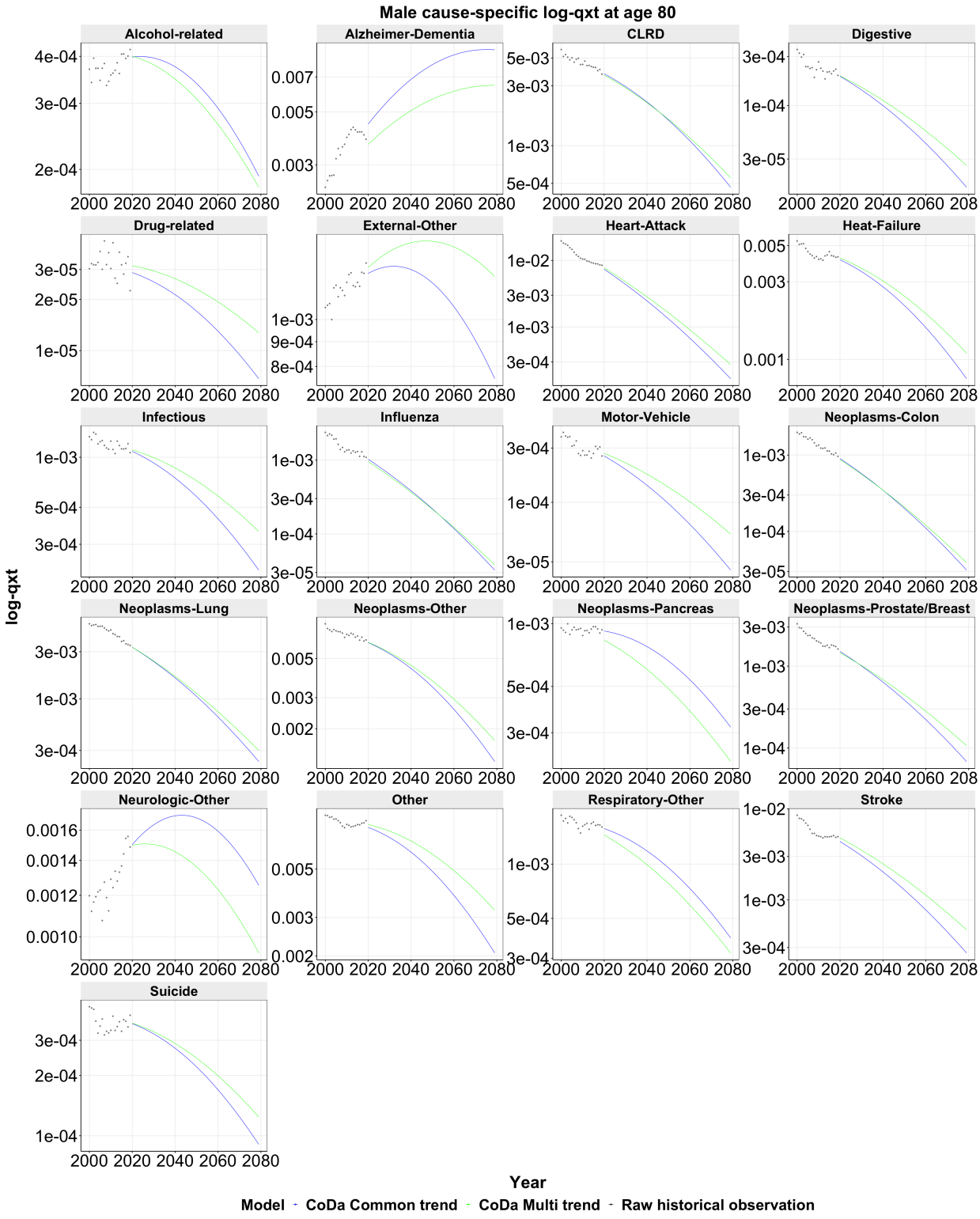


Figure 24: Male cause-specific forecast at age 80

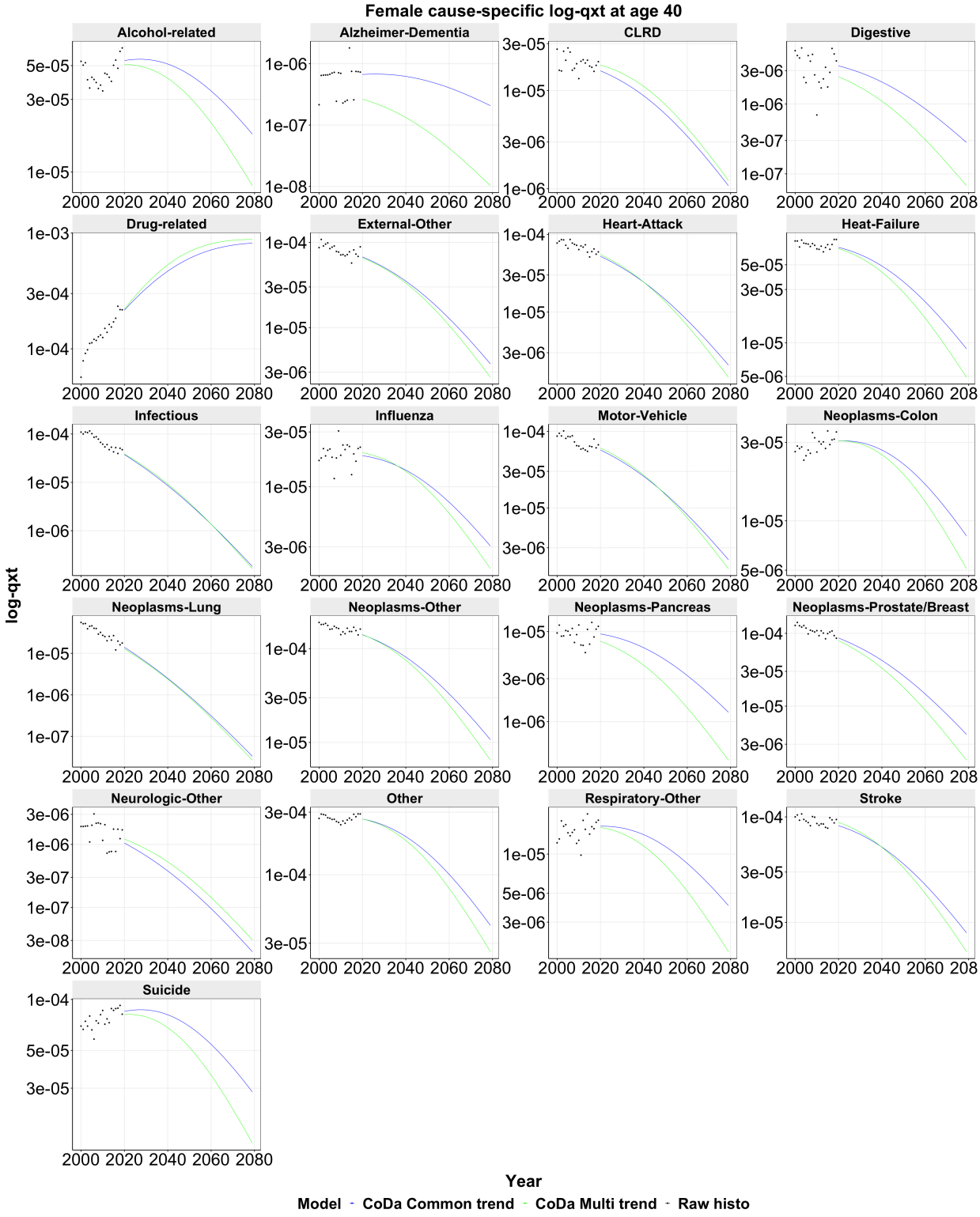


Figure 25: Female cause-specific forecast at age 40

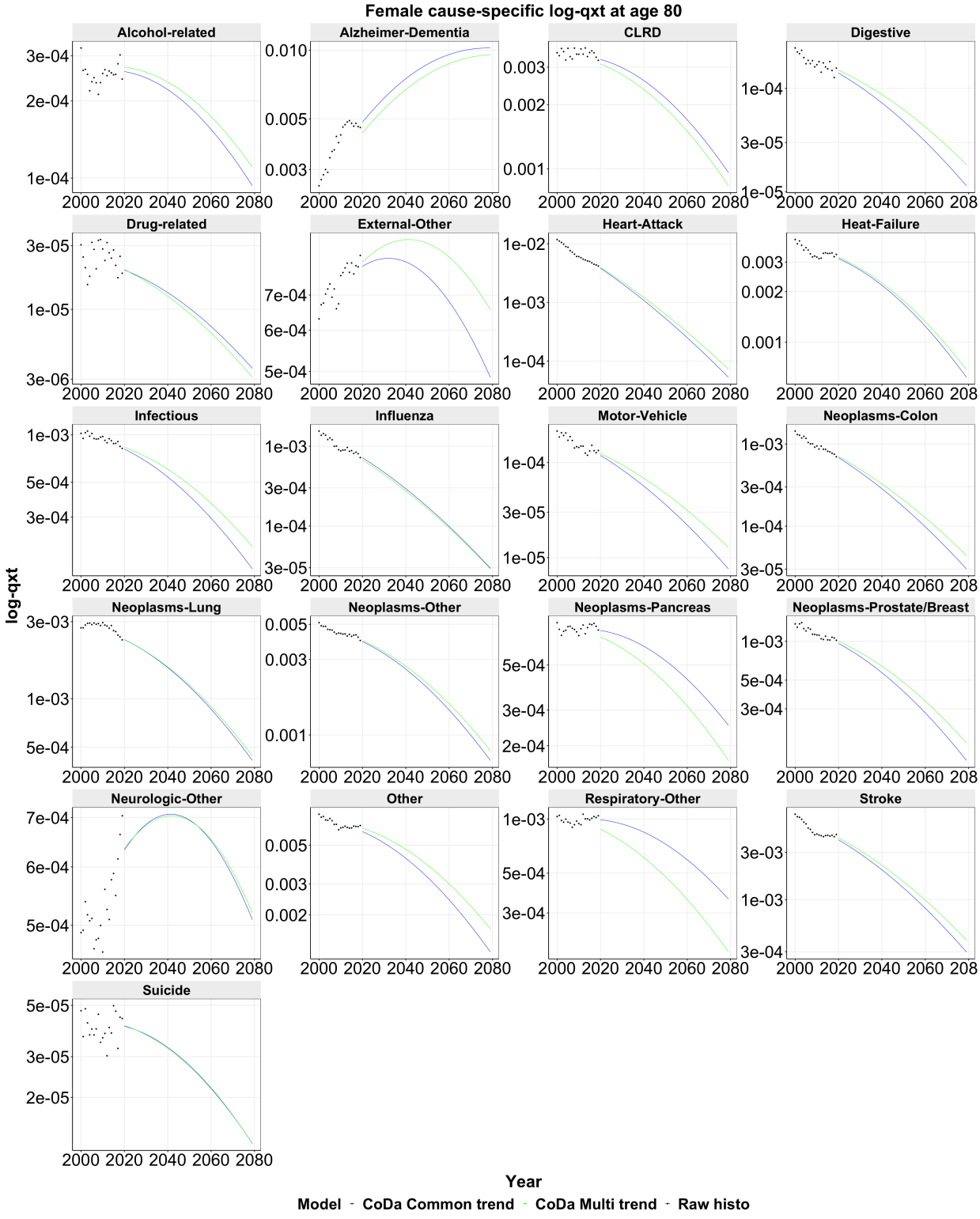


Figure 26: Female cause-specific forecast at age 80