Clustering of pathologies: application to Long-Term Care Insurance

Léonie Le Bastard

Clustering of pathologies: application to Long-Term Care Insurance

Léonie Le Bastard [1,2]

[1]Université de Lyon, Université Claude Bernard Lyon 1, Laboratoire de Sciences Actuarielle et Financière, 50 Avenue Tony Garnier, F-69007 Lyon, France
[2]SCOR SE, 5 avenue Kléber, 75795 Paris Cedex 16, France

2024/03/18

Abstract

Long-term care products cover the risk of permanent loss of autonomy. In the event of a loss of autonomy, the insurer must pay an annuity until the death or recovery of the insured. As recovery is very unlikely, long-term care product risk modellers in many countries assume that the only cause of ending annuity payments is death. Therefore, estimating the mortality of disabled insured individuals is crucial for insurers, as it significantly impacts the pricing and reserving of long-term care products. Multiple pathologies can lead to the loss of autonomy of an individual. Experience data show that the pathology responsible for the long-term care needs of an individual has a significant impact on its mortality. Therefore, accounting for the pathology is important, especially for reserving. However, the small number of data observations does not enable insurers to estimate a single mortality table for each pathology.

In this paper, we present two clustering approaches to create groups of pathologies with similar mortality rates. This allows us to aggregate the data of different pathologies and reduce the number of different mortality tables without losing too much specific information on each pathology. The first method relies on GLM trees, while the second method is a generalized K-means approach. We then show that accounting for pathologies using the clusters obtained from the proposed methods improves the predictive performance of mortality. Finally, estimating different mortality rates according to pathology allows insurers to improve reserving and to study the impact of an increase or decrease in the incidence of a specific pathology on the mortality of disabled insured individuals.

Keywords: Clustering; Pathologies; Long-Term Care Insurance; Actuarial modelling; Generalized Linear Models; GLM Trees; K-Means.

1 Introduction

Most long-term care insurance products cover the risk of loss autonomy by providing an annuity until death. Therefore, the estimation of mortality after a loss of autonomy is crucial for insurers. An overestimation of mortality would lead to an underestimation of the premium, while an underestimation of mortality could prompt individuals to underwrite the contract with another insurer due to an overestimation of the premium.

Multiple pathologies can lead to the loss of autonomy and the need for long-term care (LTC). The principal causes of such autonomy loss are dementia, cancer, neurological disease, and cardiovascular disease. Less common causes include respiratory and osteoarticular diseases, as well as accidents. As shown in Biessy (2016), pathologies have a significant impact on mortality in LTC. In this paper, the author uses a multi-state model with the states of autonomy, death and four states of LTC, combined with a continuous semi-Markov framework. The four states of LTC correspond to four different groups of pathologies: "cancer", "dementia", "neurological diseases" and "other causes". Therefore, neglecting pathology in the estimation of mortality leads to a loss of information for the insurer.

Due to the recent development of long-term care products, insurers often lack data related to the mortality of disabled insured individuals. Furthermore, information about pathologies is not always available. This scarcity of claims makes it difficult or impossible for them to estimate a specific mortality table for each pathology. Moreover, for operational reasons, insurers often prefer the simplicity of having a lower number of tables while capturing most of the variance. Clustering pathologies with similar mortality rates seems a good compromise to reduce the number of different mortality tables without losing too much specific information on each pathology.

As in Biessy (2016), we consider that mortality in LTC depends on age and duration. We are thus in a context of two-dimensional mortality laws. Therefore, this paper aims to study surface clustering methods.

The problem of curve clustering was previously addressed by Abraham et al. (2003). As the problem consists of clustering functions with an infinite dimension, Abraham et al. (2003) proposed a two-step approach. In the initial step, the dimension of each element is reduced by approximating the curve with a finite number of parameters. In this paper, the authors rely on B-splines with the following condition: the splines have to be identically distributed on the interval for each curve. In the second step, the K-means method is applied to the B-spline coefficients. Theoretically, this approach could be extended to surface clustering for pathologies. In this context, one might have to independently estimate B-spline coefficients for each element. Pathologies would then be clustered using the K-means algorithm on the basis spline coefficients. However, this method may not be suitable for our problem. Indeed, the scarcity of observations for some pathologies makes it challenging or even impossible to fit two-dimensional P-splines or B-splines for certain elements.

Clustering methods have already been used in the context of mortality modelling. Using fuzzy

clustering implemented in the R package e1071, Debón et al. (2017) clustered mortality surfaces of EU countries. In this method, each country can be associated with more than one cluster. A coefficient indicates the strength of the association between a country and each of the clusters. More recently, relying on data from the Human Mortality Database, Léger and Mazzuco (2021) used three different functional clustering methods to group countries with similar mortality trends.

Two methods, both relying on the generalized linear model (GLM) framework (Nelder and Wedderburn, 1972; McCullagh, 2019), are developed in this paper to cluster pathologies into homogeneous groups in terms of mortality. The first one, presented in Section 4.1, uses GLM trees. The second approach, presented in Section 4.2, is inspired by K-means, the most famous nonhierarchical clustering technique.

We start by describing the dataset used in this paper in Section 2 and analysing the heterogeneity induced by the plurality of pathologies. Then, Section 3 focuses on the basics of mortality modelling in the context of long-term care insurance and how generalized linear models are used for this purpose in the remainder of the paper. Clustering methods are then presented and applied to our dataset in Section 4. The performances of the clustering methods are then compared in Section 5. Using the clusters resulting from the best method according to the previous section, the benefits of accounting for pathology when modelling the mortality of disabled policyholders are discussed in Section 6. Section 7 concludes this paper by summarising the results and providing suggestions for future research.

2 Data

This section begins by introducing the dataset used for this study. We then analyse the heterogeneity of mortality between pathologies using descriptive statistics, highlighting the importance of considering this covariate in the estimation of mortality for disabled policyholders.

2.1 Presentation of the data

This study relies on data from a non-French health fund. Given that this paper focuses on mortality in LTC, only disabled policyholders are retained in the database. A total of 11,115 disabled individuals were observed from 2008 to 2016. Unlike French LTC insurance products, this health fund considers recovery from disability as being possible. The definition used for disability differs from the usual definitions in the French market, leading to the presence of transitions from disability to the autonomous state in the database. Given that this paper focuses only on mortality, recovery is treated as right censoring. Additionally, this health fund imposes a maximum lifetime benefit of five years. As a consequence, each disabled policyholder exits the disabled state after a maximum of 5 years, and health status is not observed beyond this period. This maximum lifetime benefit implies right censoring, and no observations are available for durations exceeding 5 years. Information regarding the pathology responsible for the loss of autonomy of each disabled policyholder is available.

The detailed diagnosis is provided by the health fund, along with a broader pathology group at a more macro level. Examples of pathology groups include "Respiratory diseases", "Cardiovascular diseases" and "Dementia". The health fund distinguishes 14 groups of pathologies. Rare pathologies are grouped as "Other". A total of 5,000 deaths of disabled insured individuals were observed during the observation period.

Table 1 summarises the number of claims and observed deaths per pathology. In this health fund, many disabled insured individuals have multiple pathologies identified as the cause of the loss of autonomy. The most prevalent causes of claims include dementia, cancer, cardiovascular diseases, and accidents.

Pathology	**Censoring**	**Death**	**Total**
Accident	740	295	1,035
Cancer	504	1,501	2,005
Cardiovascular disease	752	494	1,246
Dementia	1,235	907	2,142
Endocrine disease	28	28	56
Gastrointestinal disease	26	16	42
Infectious disease	31	15	46
Neurological disease	427	174	601
Osteoarticular	136	63	199
Other	360	88	448
Psychiatric disease & depression	198	132	330
Respiratory disease	55	67	122
Several	1,601	1,159	2,760
Urological or kydney disorder	32	51	83

Table 1: Number of observed insured individuals and causes of exit for each pathology

Although the original data are at the individual level, the models presented in the paper use aggregated data. The observations are grouped by attained age, duration, gender, and pathology by defining a discretization grid for age and duration. Age and duration intervals are split on an annual basis. However, due to significant variations in mortality in the early onset of disability, the first year following the loss of autonomy is commonly partitioned on a monthly basis. As a result, the database used in the remainder of the paper contains one line for each combination of these four variables. For each of these lines, the central exposure and count of deaths are calculated using the original individual database.

Notation: x_i denotes the i^{th} split point of the discretization grid on the age interval, and t_j denotes the j^{th} split point on the duration interval.

The central exposure associated with the combination of age x_i, duration t_j, gender g and pathology

p is the sum of the individual central exposures of disabled policyholders with pathology p and gender g, for an attained age between x_i and x_{i+1} and a duration between t_j and t_{j+1}.

2.2 Heterogeneity of mortality of disabled insured individuals between pathologies

Biessy (2016) shows the significant impact of pathologies on mortality within the French LTC insurance market. Our data exhibit the same phenomenon. The impact of pathologies on mortality can be illustrated by calculating the ratios of the actual counts of deaths over the expected counts under the null hypothesis H_0: "Given the gender, age and duration since the loss of autonomy, the mortality rates are equal for all pathologies".

Assuming that mortality rates are independent of the pathology responsible for the loss of autonomy, crude mortality rates are estimated on the entire database of disabled policyholders for each gender, without accounting for the pathology. The expected counts of deaths for each pathology, attained age, and duration are then computed using the central exposures and crude mortality rates of disabled policyholders.

The ratios of actual over expected counts of deaths for each pathology and gender are plotted in Figure 1. Under the assumption that pathology has no impact on mortality in LTC, the observed counts of deaths should be close to the expected counts. Therefore, the actual-to-expected ratios should be close to 1. With a significance level of $\alpha = 5\%$ and under the Poisson assumption of the number of deaths, the tolerance intervals of the ratios of actual over expected values are constructed. The purpose of these intervals is to account for the variance in the random variable of the observed counts of deaths. Considering pathology p, if the ratio is above the tolerance interval, we can conclude that the mortality observed with pathology p is significantly underestimated by the common mortality law aggregating all pathologies. If the ratio is below this interval, the mortality is significantly overestimated. Figure 1 shows that using common mortality rates for all pathologies leads to a significant underestimation of the mortality of disabled insured individuals with cancer for both males and females. Moreover, using a unique mortality law for all pathologies results in an overestimation of mortality of disabled policyholders with cardiovascular disease, dementia, neurological disease, psychiatric disease, as well as those disabled because of an accident. Further details about statistical hypothesis testing using the metric of actual over expected values are provided in Section 5.2.1.

This finding highlights the importance of considering pathology when estimating the mortality of disabled policyholders in the context of modelling LTC insurance products. However, datasets containing information about the pathology leading to the need for long-term care are scarce, and some pathologies are also poorly represented in portfolios. Table 1 shows that fewer than 100 individuals were observed by the health fund during the study period for the following pathologies:

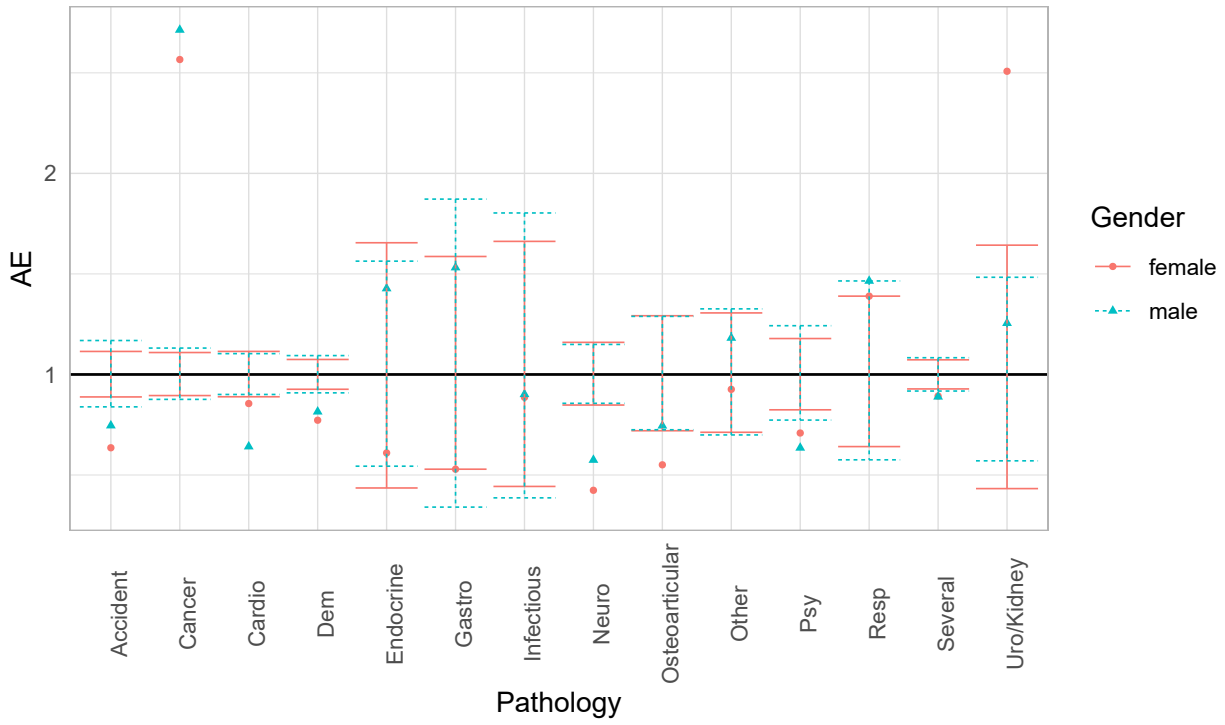

Figure 1: Actual-to-expected ratios obtained by considering a common mortality law for all pathologies

endocrine diseases, gastrointestinal diseases, infectious diseases, and urological or kidney disorders. Therefore, while accounting for pathologies is crucial, the lack of data for some of them makes estimating a specific mortality table for each pathology challenging. Moreover, for the sake of simplicity, insurers prefer to maintain a reasonable number of distinct mortality tables for disabled policyholders. Therefore, clustering pathologies with similar mortality rates is an effective means of enhancing mortality modelling without introducing excessive complexity.

3 The fundamentals of mortality modelling using the GLMs

This section focuses on the common tools used to model and estimate mortality of disabled policyholders in long-term care insurance products. We first describe the tools for survival analysis and the common assumptions used for such products in Section 3.1. Two different methods for clustering pathologies are presented and tested in this paper. Since both approaches rely on generalized linear models (GLMs), we describe the fundamentals of GLMs as applied to mortality modelling in Section 3.2.

3.1 Modelling the mortality of disabled policyholders

In the survival model theory, mortality is often described by the force of mortality μ, also called mortality intensity. Numerous papers on LTC, such as Czado and Rudolph (2002), Biessy (2017)

and Biessy (2019), show that the force of mortality of disabled policyholders depends on the time that has elapsed since the loss of autonomy. Semi-Markov models, initially introduced in the context of disability modelling by Janssen (1966), followed by Hoem (1972), assume that the intensities of the process depend not only on the current state but also on the time spent in this state. As in more recent papers, such as Pitacco (2014) for disability modelling and Soetewey et al. (2022), Fuino and Wagner (2018), and Xuanyuan Shihao and Shiang Xuanyuan (2023) for long-term care insurance modelling, we assume the following:

Assumption. 3.1

Mortality rates depend on attained age and time spent since the loss of autonomy.

Let $\mathcal{X}_x$ denote the health status of a disabled policyholder at age x. Let A, LTC and $Death$ denote the autonomous, disabled and death states, respectively. The force of mortality $\mu_{x,t}$ is given by

$$\mu_{x,t} = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = D | \mathcal{X}_x = LTC, \mathcal{X}_{(x-t)} = LTC, \mathcal{X}_{(x-t)^-} = A\right)}{h} \tag{1}$$

The force of mortality $\mu_{x,t}$ is commonly assumed to be piecewise constant by age and duration, as in Guibert and Planchet (2018) and Soetewey et al. (2022). This assumption is reasonable for appropriate steps of age and duration. In the remainder of the paper, we assume the following:

Assumption. 3.2

Mortality rates are piecewise constant by age and duration on a yearly basis, except for the first year following the loss of autonomy, in which mortality rates are piecewise constant on a monthly basis on the duration axis.

The split points of the age and duration intervals are denoted as $(x_i)_{i \in \{1,\dots,M_x\}}$ and $(t_j)_{j \in \{1,\dots,M_t\}}$, respectively.

Mortality can also be described by q_{x_i,t_j}, denoting the probability of an individual of integer age x_i who has been disabled for t_j years dying before reaching duration t_{j+1} or age x_{i+1}. μ_{x_i,t_j} and q_{x_i,t_j} are related by the following equation:

$$\begin{aligned} q_{x_i,t_j} &= 1 - \exp\left(-\int_0^{\min(t_{k+1}-t_k, x_{i+1}-x_i)} \mu_{x_i+u,t_j+u} \mathrm{d}u\right) \\ &= 1 - \exp\left(-\mu_{x_i,t_j} \times \min\left(t_{k+1} - t_k, x_{i+1} - x_i\right)\right). \end{aligned} \tag{2}$$

As men and women are not equal when it comes to the risk of mortality and loss of autonomy, the mortality rates of disabled policyholders depend on gender g.

We denote by $\mu^{g,p}_{x_i,t_j}$ and $q^{g,p}_{x_i,t_j}$ the force of mortality and the probability of death associated with gender g and pathology p for age $x, x_i \leqslant x < x_{i+1}$ and duration $t, t_j \leqslant t < x_{j+1}$, respectively.

3.2 Basics of GLMs

Both methods used in this paper to cluster pathologies rely on generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh, 2019). GLMs are a generalization of linear models that allow the probability distribution of the response variable Y to be chosen from the exponential family introduced in Barndorff (1978). Examples of distributions from this family are Normal, Exponential, Log-Normal, Gamma, Binomial, Poisson, and Inverse Gaussian. Let X denote the covariates. The conditional mean of the response variable given X, is expressed as a function of a linear combination of X through the following expression:

$$\mathbb{E}[Y|X] = G^{-1}(X\beta), \tag{3}$$

where β is a coefficient vector, $G(\cdot)$ is the link function, and $X\beta$ is the linear predictor.

The GLM depends on three elements:

- the probability distribution,
- the link function $G(\cdot)$,
- the linear predictor $\eta = X\beta$.

The optimal coefficient vector β is obtained by maximizing the log-likelihood.

The cross effects of a qualitative covariate with another covariate (qualitative or quantitative) can be included in the linear predictor to allow group-specific coefficients in the model. In this case, some elements of the vector of coefficients β may differ depending on the value of the qualitative covariate involved in the cross effect.

In some situations, it may also be useful to add an additional covariate to the linear predictor, with a fixed coefficient that does not have to be estimated. This term is called an offset. In this case, Equation 3 becomes

$$G(\mathbb{E}[Y|X]) = X\beta + \text{offset}. \tag{4}$$

In the context of mortality modelling, the response variable is the number of deaths. The common probability distributions used in this context are Poisson and Binomial distributions (Hunt and Blake, 2021).

Let $D^{g,p}_{x_i,t_j}$ denote the number of deaths occurring at age x between x_i and x_{i+1} with a duration in claim t between t_j and t_{j+1} for gender g and pathology p.

Assuming a binomial distribution of the count of deaths, $\mathbb{E}[D^{g,p}_{x_i,t_j}] = {}^0e^{g,p}_{x_i,t_j} q^{g,p}_{x_i,t_j}$, where ${}^0e^{g,p}_{x_i,t_j}$

denotes the initial number of disabled live policyholders with pathology p at age x_i with a duration of claim t_j. Assuming a Poisson distribution of the count of deaths, $\mathbb{E}[D^{g,p}_{x_i,t_j}] = e^{g,p}_{x_i,t_j}\mu^{g,p}_{x_i,t_j}$, where $e^{g,p}_{x_i,t_j}$ denotes the central exposure to risk (average number of disabled policyholders with pathology p and gender g with attained age between x_i and x_{i+1}, for a claim duration between t_j and t_{j+1}).

While central exposures are widely available and easily estimated from a database, initial exposures are more complex to estimate because of censoring and truncating. However, initial exposures can be approximated from central exposures. Both distributions were tested in this study, but we retained the Poisson distribution due to its better predictive performance in terms of the log-likelihood values.

In GLMs, the link function is chosen by the user and determines the link between the regressor and the estimated value of the response. However, as explained in Myers and Montgomery (1997), each distribution has a special link function called the canonical link function, which has the nice mathematical properties described in Nelder and Wedderburn (1972). The canonical function of the Poisson GLM is the log function. In this paper, we work with the canonical link function because of its useful properties.

With the number of deaths as the response variable, and using the Poisson assumption combined with the log link function, we have the following equation:

$$\begin{aligned} \log(\mathbb{E}[D^{g,p}_{x,t}] &= \log(e^{g,p}_{x,t}\mu^{g,p}_{x,t}) \\ &= \underbrace{\log(e^{g,p}_{x,t})}_{\text{offset}} + \log(\mu^{g,p}_{x,t}). \end{aligned} \tag{5}$$

Therefore, using the Poisson regression to model the counts of deaths, an offset accounting for the central exposure is added to the linear predictor.

The covariates considered in this paper are the gender, age and duration. In what follows, we use a formula assuming a quadratic impact of age, and a cubic impact of duration. As mortality in LTC greatly differs for males and females, a cross effect of gender with both age and duration is considered in the GLM formula.

4 Clustering methods and initial results

The first clustering approach, which rely on GLM Trees, is described in Subsection 4.1. The second method proposed in this paper is inspired by the well known K-means algorithm. This method will be further explained in Subsection 4.2.

4.1 First method: GLM trees

As can be seen from its name, this method introduced by Achim Zeileis and Hornik (2008) combines generalized linear models (GLMs) with decision tree models. The decision tree algorithm is a unsupervised learning method, that builds a tree of clusters. The root cluster contains all the observations, and each node contains a subset of the data of its parent. The tree successively splits each node according to a specific metric. If all the observations in the node are considered to belong to the same group according to the metric, then the node is not split further. The final leaves indicate the final label of each observation.

The GLM tree method is as follows. As in Section 3.2, let Y denote the variable to be explained, and let X denote the covariates included in the linear predictor to explain Y. Let Z denote the covariates considered for splitting the tree and called the split variables. A variable can be used both as a split variable and as a covariate in the linear predictor in a GLM, at the same time. As with basic decision trees, the root node contains all the observations. At each step, for each node of the current level of the tree, the algorithm starts by estimating the GLM parameters using all the observations from that node. Thereafter, the stability of the parameters is tested over the split variables. The node is then split according to the split variable with the highest parameter instability. These steps are repeated until the parameter stability threshold is met or the tree has reached a maximum depth chosen in advance by the user. According to Achim Zeileis and Hornik (2008), one of the main benefits of this method is the use of the same objective function for both parameter estimation and partitioning. While the formula of the GLM fitted in each node implies linear relationships between Y and the covariates X, recursive partitioning allows for a nonlinear relationship by allowing cuts of quantitative variables at some identified optimal split points.

The formula of the GLM tree is written as follows:

$$Y \sim f(X)|Z,$$

where $f(X)$ describes the form of the linear predictor.

In this paper, we focus on the clustering of pathologies. For the sake of simplicity, insurers prefer to have identical clusters of pathologies for both genders. Therefore, the only variable considered as a split variable is pathology. The covariates considered as explanatory variables in the GLM are gender, attained age of the insured, and duration since the loss of autonomy. GLM trees can be fitted using a recursive partitioning algorithm implemented in **R** (R Core Team, 2020) within the **partykit** library. A detailed description of the library is provided in Hothorn and Zeileis (2015).

The formula used for our application is:

$$\text{Formula: } Count \sim Age \times Gender + Age^2 \times Gender + Duration \times Gender + Duration^2 \times Gender + Duration^3 \times Gender|Pathology.$$

Assuming a Poisson distribution of the counts of deaths and using the log link function, an offset accounting for the central exposure in each cluster is added to the GLM, as presented in Equation

4. Let $\kappa(p)$ denote the assigned cluster of pathology p; the offset for cluster k is given by

$$\text{Offset}_{x,t}^{g,k} = \log\left(\sum_{p} \mathbb{1}\{\kappa(p) = k\} e_{x,t}^{g,p}\right)$$

Figure 2 shows the results of the GLM tree clustering. The analysis of the actual-to-expected ratios in Section 2.2 shows that the mortality of disabled policyholders who lose their autonomy because of cancer greatly differs from the mortality associated to other pathologies. This difference in mortality is detected by the GLM tree, which sorts cancer into its own cluster. Let us analyse the recursive partitioning in parallel with the results of the actual-to-expected ratios plotted in Figure 1. We use "global disabled mortality" to denote the mortality of the overall portfolio of disabled policyholders without accounting for the pathology. The GLM tree seems to first separate the pathologies whose observed mortalities are underestimated by the global disabled mortality, from the other pathologies. These include cancer; respiratory disease; urological and kidney disorders; and the "other" group, corresponding to all pathologies not listed among the 14 studied pathologies.

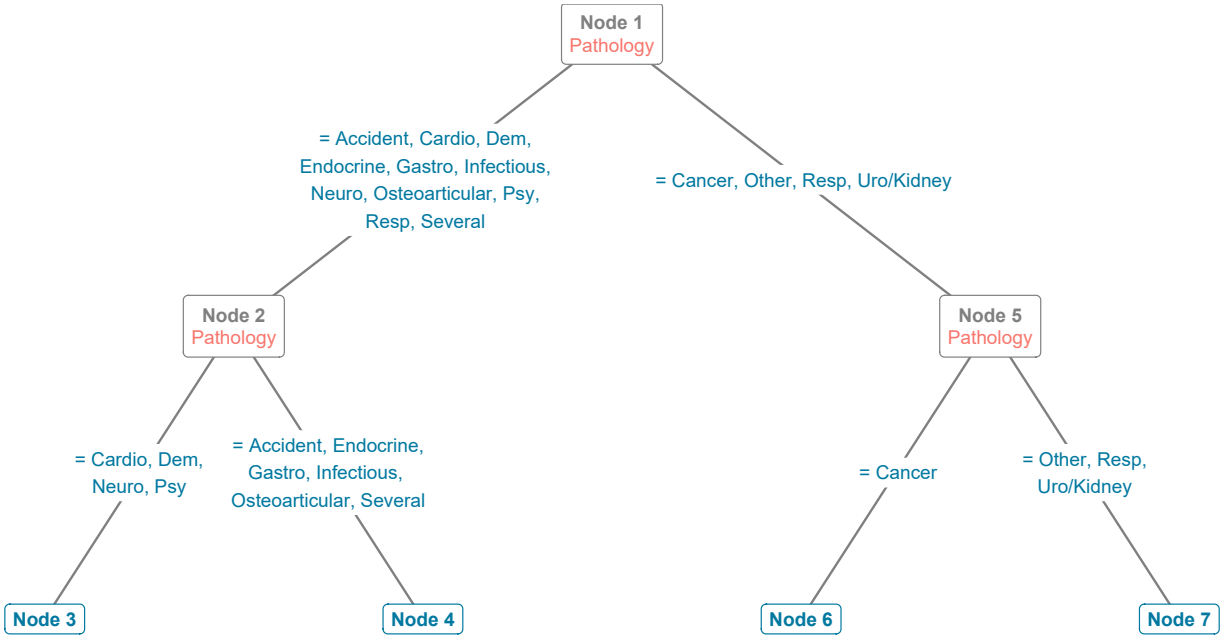

Figure 2: Resulting tree obtained by the GLM tree method

4.2 Second method: Generalized K-means

K-means is one of the most widely used unsupervised clustering algorithms. In this section, we propose a K-means approach to cluster pathologies with similar mortality rates. Section 4.2.1 starts by introducing the basics of the K-means algorithm and the chosen features required to apply it to our problem of surface clustering. The first step of the K-means approach depends on a random distribution of the pathologies among the K groups. Section 4.2.2 studies the impact of this random

initialization on the composition of the final clusters. Since this impact seems rather significant in our application, a similarity matrix is built by running the K-means algorithm many times. Two approaches presented in Section 4.2.3 are then subsequently tested to cluster pathologies based on this similarity matrix.

4.2.1 Generalized K-means for pathology clustering

After fixing the desired number of final clusters K, the goal of clustering is to make the elements within the clusters as similar as possible while maximizing the dissimilarity between elements of different groups. K-means clustering requires the choice of a distance or similarity measure between single elements and clusters. Let a be an element and C denote a cluster. If the elements are points from an Euclidian space $\mathbb{R}^n$, then the common distance measure $dist(a, C)$ is the Euclidian distance,

$$dist(a, C) = \sum_{i=1}^{n} (a_i - C_i)^2.$$

Any other similarity or dissimilarity measure can be used to adapt the K-means algorithm depending on the context and the problem. In this study, each pathology p is represented by an element a^p containing information about the observed counts of deaths and central exposures by attained age, duration and gender. In this work, mortality associated with each cluster is estimated using GLMs. Therefore, our method combines GLMs with K-means. The use of K-means in a GLM framework has already been addressed in several papers as in Zhang and Lin (2021) and Abraham et al. (2003). The latter proposed a two-step approach for unsupervised curve clustering. After fitting B-splines to each element, the similarity between curves is assessed by measuring the distance between the vectors of the spline coefficients. This approach is based on the idea that curves with similar B-spline coefficients are close. However, this approach cannot be used in our work because of the rather limited number of observations for certain pathologies, preventing a robust estimation of their associated spline coefficients.

Let $C^k, k \in \{1, \dots, K\}$ denote K distinct clusters. Each cluster can be represented by its mortality surface, which represents the common mortality of all the pathologies within this cluster. The similarity $S(a^p, C^l)$ between an element a^p and a cluster C^k is assessed by the Poisson log-likelihood of the observed counts of deaths from pathology p assuming the mortality rates associated with cluster C^k. The distance measure is given by the following equation:

$$\begin{aligned} dist(a^p, C^k) &= -S(a^p, C^k) \\ &= -\left(\sum_{g=1}^{2} \sum_{i=1}^{M_x} \sum_{j=1}^{M_t} -\mu_{x_i,t_j}^{g,k} e_{x_i,t_j}^{g,p} + d_{x_i,t_j}^{g,p} \log\left(\mu_{x_i,t_j}^{g,k} e_{x_i,t_j}^{g,p}\right) \right), \end{aligned} \tag{6}$$

where:

- M_x is the number of subdivisions in the age interval,

- M_t is the number of subdivisions in the duration interval,
- $\mu^{g,k}_{x_i,t_j}$ is the common mortality intensity of pathologies in cluster C^k, and
- $e^{g,p}_{x_i,t_j}$ is the central exposure of disabled policyholders of gender g with pathology p.

Remark: As in Section 4.1, we assume similar clusters of pathologies for males and females.

The steps of the adapted K-means clustering algorithm are as follows:

Algorithm. 4.1: K-means clustering algorithm

1. Randomly assign a cluster from 1 to K to each pathology in the dataset, corresponding to the initial cluster of the pathology,
2. Iterate the following until the distribution of pathologies among the clusters stops changing:
   (a) Fit the GLM for each cluster without accounting for the pathology. The central exposures and counts of deaths from pathologies within the same cluster are added.
   (b) Assign each pathology to the cluster with the closest mortality surface, using the distance measure $dist(\cdot,\cdot)$ from Equation 6.

Since K-means requires fixing the number of clusters in advance, our method is combined with the generalized information criterion (GIC) to select the best hyperparameter K when this value is unknown. As recommended in Zhang and Lin (2021), K is chosen to minimize the Bayesian information criterion (BIC). The choice of K is further discussed in Section 5.1.

4.2.2 Impact of the random initialization and construction of a distance matrix

The first step of the adapted K-means algorithm is to randomly assign a cluster to each of the M_P pathologies. Let us analyse the impact of this random initialization on the composition of the final groups of diseases for a chosen number of final clusters $K = 4$. We then propose a method to construct a matrix of distances between pathologies based on their assignments to the K groups for a large number of random initializations.

The K-means algorithm is run successively with 1,000 random initializations. The distribution of the assigned cluster for each pathology is plotted in Figure 3.

After building the 4 groups and fitting a GLM to estimate the mortality surfaces associated with them, the groups were ranked by the level of mortality at age 70 for the first duration. This arbitrary choice is used to label the clusters in each iteration. As shown in Figure 3, cancer seems to always be assigned to the cluster 4, which has the highest mortality rate at 70 years old immediately after the loss of autonomy. In contrast, other pathologies are rarely assigned to this group. This finding shows that cancer differs from other pathologies in particular by its high mortality rate for low durations. Interpretation is more difficult for other pathologies since random initialization seems to have a greater impact on their final assigned clusters.

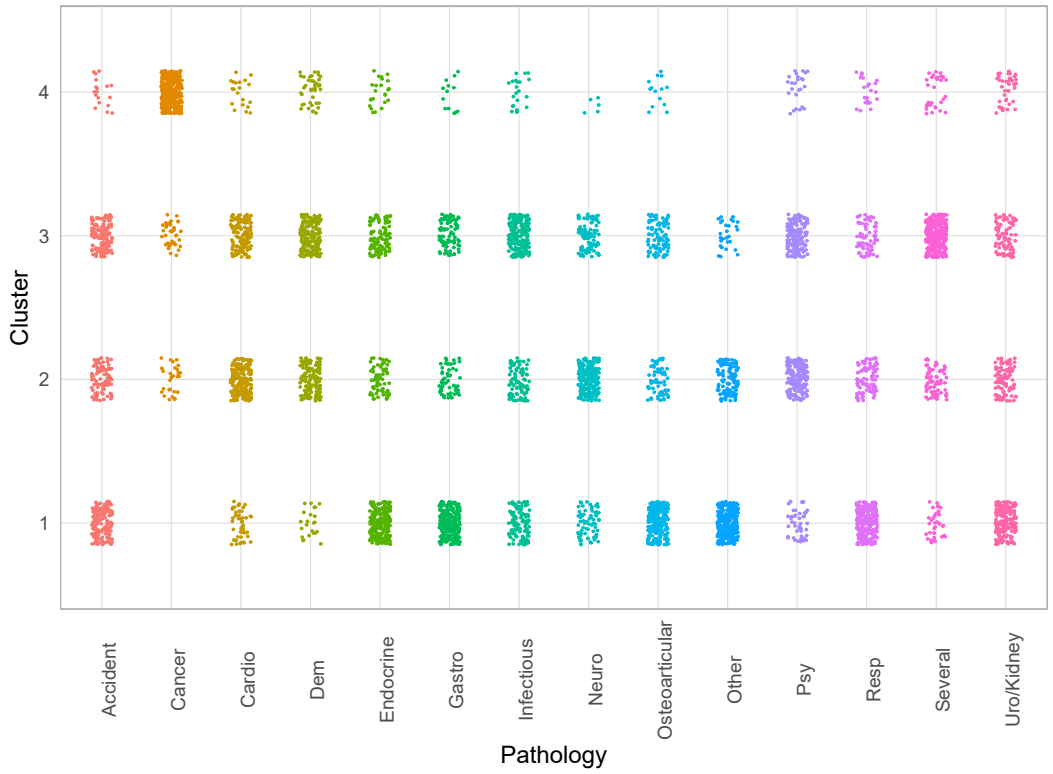

Figure 3: Impact of the random initialization on the cluster assigned to each pathology

We therefore need a metric to assess the similarity between pathologies. This can be estimated by answering the following question: how often are two pathologies assigned to the same cluster? A similarity matrix $\mathcal{S} \in \mathcal{M}_P(\mathbb{R})$ is built, where $\mathcal{S}_{p,\tilde{p}} \in [0;1]$ denotes the similarity between pathologies p and $\tilde{p}$.

Let R denote the number of random initializations. Let $\kappa^r(p) \in \{1,\dots,K\}$ denote the assigned cluster of pathology p for the r^{th} random initialization; then $\mathcal{S}_{p,\tilde{p}}$ is estimated by

$$\mathcal{S}_{p,\tilde{p}} = \frac{\sum_{r=1}^{R} \mathbb{1}\{\kappa^r(p) = \kappa^r(\tilde{p})\}}{R}. \tag{7}$$

Note: $\mathcal{S}$ is symmetric by construction, and $\mathcal{S}_{p,p} = 1$.

The similarity matrix $\mathcal{S}$ associated with the $R = 1{,}000$ random initializations plotted in Figure 3 is given by Table 2. Values above 0.6 in the upper diagonal are highlighted in yellow for values between 0.6 and 0.7, orange for values between 0.7 and 0.8 and red for values above 0.8. The two pathologies the most often assigned to the same group are dementia and psychiatric diseases. Then, urological and kidney diseases are assigned to the same cluster as respiratory diseases in 83% of the 1,000 initializations. With all similarity values between 0.6 and 0.8, we can see in yellow and orange

that endocrine, gastrointestinal, osteoarticular, infectious and respiratory diseases seem similar in terms of mortality when they are responsible for the loss of autonomy. Other similarity values are more difficult to interpret. Therefore, statistical methods are needed to construct optimal clusters based on this similarity matrix. Two approaches are presented and tested in Section 4.2.3 to address this problem.

	Accident	Cancer	Cardio	Dem	Endocrine	Gastro	Infectious	Neuro	Osteo-articular	Other	Psy	Resp	Several	Uro Kidney
Accident	1.00	0.00	0.22	0.34	0.55	0.49	0.62	0.38	0.57	0.38	0.32	0.37	0.40	0.24
Cancer	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.02
Cardio	0.22	0.00	1.00	0.55	0.10	0.01	0.12	0.57	0.08	0.20	0.46	0.18	0.41	0.18
Dem	0.34	0.00	0.55	1.00	0.10	0.07	0.19	0.62	0.10	0.10	0.84	0.06	0.39	0.03
Endocrine	0.55	0.00	0.10	0.10	1.00	0.78	0.65	0.04	0.78	0.55	0.06	0.71	0.46	0.59
Gastro	0.49	0.00	0.01	0.07	0.78	1.00	0.62	0.06	0.79	0.45	0.04	0.64	0.32	0.53
Infectious	0.62	0.00	0.12	0.19	0.65	0.62	1.00	0.13	0.63	0.30	0.14	0.52	0.60	0.49
Neuro	0.38	0.00	0.57	0.62	0.04	0.06	0.13	1.00	0.11	0.09	0.64	0.04	0.22	0.02
Osteo-articular	0.57	0.00	0.08	0.10	0.78	0.79	0.63	0.11	1.00	0.46	0.10	0.60	0.45	0.47
Other	0.38	0.13	0.20	0.10	0.55	0.45	0.30	0.09	0.46	1.00	0.07	0.69	0.20	0.56
Psy	0.32	0.00	0.46	0.84	0.06	0.04	0.14	0.64	0.10	0.07	1.00	0.01	0.28	0.00
Resp	0.37	0.00	0.18	0.06	0.71	0.64	0.52	0.04	0.60	0.69	0.01	1.00	0.40	0.83
Several	0.40	0.00	0.41	0.39	0.46	0.32	0.60	0.22	0.45	0.20	0.28	0.40	1.00	0.40
Uro Kidney	0.24	0.02	0.18	0.03	0.59	0.53	0.49	0.02	0.47	0.56	0.00	0.83	0.40	1.00

Table 2: Similarity matrix for $K = 4$

4.2.3 Clustering based on the similarity matrix

Using the similarity matrix $\mathcal{S}$ built with R iterations of generalized K-means for a fixed number of final clusters K, the goal of this section is to construct K clusters of pathologies. This can be seen as an optimal permutation problem (OPP), introduced in Morone (2022). In this paper, the authors proposes a theoretical framework to find an optimal permutation of the rows and columns of the matrix to obtain a new matrix as close as possible to a desired clustered form. Indeed, if there exists a permutation matrix P such that $P^t\mathcal{S}P$ is a block matrix composed of K blocks, then each block defines a cluster. In this hypothetical situation, the similarity between pathologies of seperate clusters is zero. A clear structure of clusters where all similarity values between two pathologies of seperate groups are zero is an extreme situation. In most cases, the goal is to maximize values within the blocks while minimizing values outside the blocks. One disadvantage of this method is the need to set the number of elements within each cluster in advance in addition to the number of clusters K, leading to less flexibility in clustering. Therefore, we propose two other approaches to cluster pathologies based on the similarity matrix $\mathcal{S}$ without setting the number of elements within each group in advance. While both methods are bottom-up hierarchical clustering algorithms, as described in Algorithm 4.2, they differ in how they assess the distances between clusters.

Algorithm. 4.2: Bottom-Up hierarchical clustering algorithm

1. Assign each object to its own cluster.
2. Iterate the following until there is a single cluster:
   (a) Aggregate the two clusters with the highest similarity or, equivalently, the smallest distance;
   (b) Update the distances or similarity values between all clusters using a chosen formula.

Starting with M_P elements, the number of clusters after the l^{th} iteration is $M_P - l$. The clusters constructed at each step are called "temporary clusters" in what follows. Let $(c^k_{(l)})_{k\in\{1,\dots,M_P-l\}}$ denote the set of $M_P - l$ temporary clusters at the end of the l^{th} iteration. The difference between the two approaches lies in step 2b of Algorithm 4.2.

First method for updating the similarity between clusters

In the first approach, the computation of the distance between two temporary clusters relies on conditional probability theory. At step l, the similarity between $c^m_{(l)}$ and $c^{\tilde{m}}_{(l)}$ is assessed by the probability that these two temporary clusters are subsets of the same final cluster obtained with K-means, conditional on the following event:

"For all temporary clusters $c^m_{(l)}$ of step l, there exists a final cluster $C^k, k \in \{1,\dots,K\}$ such that all pathologies in $c^m_{(i)}$ are in C^k". This event is mathematically written as

$$\forall m \in \{1,\dots,M_P - l\}, \exists k \in \{1,\dots,K\}, c^m_{(l)} \subset C^k.$$

Therefore, the similarity between $c_{(l)}^{m}$ and $c_{(l)}^{\tilde{m}}$ after l iterations is

$$\mathbb{P}\left(\exists k \in \{1,\dots,K\}, \left\{c_{(l)}^{m} \cup c_{(l)}^{\tilde{m}}\right\} \subset C^{k} \middle| \left\{\forall m \in \{1,\dots,M_P - l\}, \exists k \in \{1,\dots,K\}, c_{(l)}^{m} \subset C^{k}\right\}\right) \quad (8)$$

Figure 4 shows the construction of the groups using this clustering approach. Each pathology has its own cluster at the beginning, as illustrated on the left. As $K = 4$ and $M_P = 14$, we do not need to go further than the 10^{th} iteration of Algorithm 4.2. The numbers in the intermediate nodes represent the iterations of the algorithm at which this aggregation occurred.

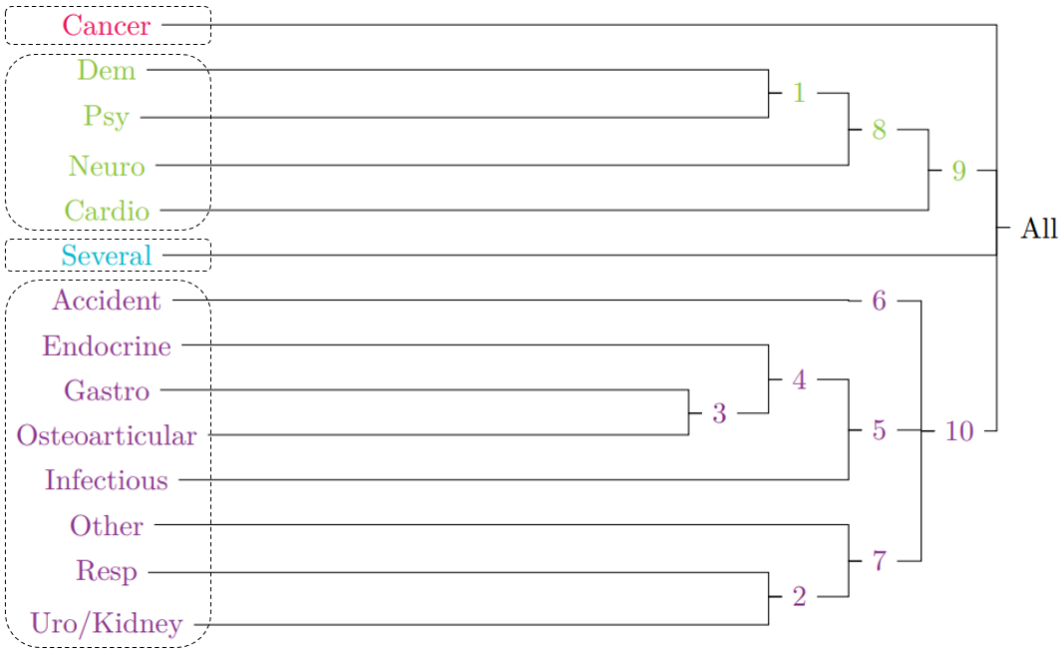

Figure 4: Dendrogram obtained with the conditional probability approach ($K = 4$, Poisson GLM, Formula $= x * Gender + I(x^2) * Gender + t * Gender + I(t^2) * Gender + I(t^3) * Gender$")

Despite having the same number of final clusters as the GLM tree approach, this method yields a different distribution of the pathologies within the groups. As anticipated in Section 4.2.3, the algorithm first groups dementia and psychiatric diseases. In the second iteration, respiratory, urological and kidney diseases are aggregated. As shown by analysis of the similarity matrix given in Table 2, cancer seems to greatly differ from other pathologies and therefore has its own cluster. The final clusters are circled using different colours in Figure 4. Despite being associated with infectious diseases in 60% of the K-means iterations, the group "Several", corresponding to the case where the disabled policyholder has multiple diseases, has its own final cluster.

Since the resulting groups are different from those obtained with the GLM tree method, the predictive performances of these two approaches will be further studied in Section 5.2.

Second method of updating the similarity between clusters

In the second approach, the distance between two temporary clusters $c_{(l)}^{m}$ and $c_{(l)}^{\tilde{m}}$ is assessed by the maximum distance between an element of $c_{(l)}^{m}$ and an element of $c_{(l)}^{\tilde{m}}$. This method is implemented in **R** within the **stats** library. The function **hclust()** implements hierarchical clustering algorithms based on a set of dissimilarities. The similarity matrix constructed in Section 4.2.2 is transformed into a dissimilarity or distance matrix $\mathcal{D} = 1 - \mathcal{S}$. While several agglomeration methods are provided in this function, corresponding to different ways of recomputing the distances at each step, the complete linkage method assesses the distances between clusters according to the maximum distance between two elements of these clusters, as is wanted here. More details about the **hclust()** function are provided in RDocumentation (2023). The resulting dendrogram, obtained with $K = 4$, is plotted in Figure 5. As anticipated in Section 4.2.2 by analysing the values of the similarity matrix $\mathcal{S}$, cancer has its own cluster because of the specificity of its mortality surface. Moreover, we are not surprised to see that psychiatric diseases and dementia are in the same group, and that urological, kidney and respiratory diseases are all included in the same cluster. While endocrine, gastrointestinal, osteoarticular, and infectious diseases are grouped in the same cluster, respiratory diseases that also look similar to endocrine diseases are grouped with urological and kidney disorders instead. This can be explained by a greater similarity with urological and kidney disorders and low similarity between respiratory diseases and all other pathologies associated with endocrine diseases.

This approach for updating the similarity measure at each step of the Algorithm 4.2 yields the same final groups as does the GLM tree approach.

5 Choice of the number of clusters in the generalized K-means methods and comparison between all clustering approaches

In this section, we compare the performances of the three clustering methods. We start by discussing the choice of the number of clusters K in the two generalized K-means approaches. Many insurance and reinsurance companies rely on expert judgments to cluster pathologies instead of using clustering algorithms. How would an expert aggregate the 14 pathologies examined in this paper to create 4 homogeneous groups in terms of mortality? We asked an epidemiologist working for the reinsurer SCOR to construct clusters of pathologies with similar mortality rates, based on her experience and expert judgment. We want to compare the performance of the model using clusters produced by expert judgment to that of the models using groups from the three clustering methods presented in this paper. After selecting the optimal hyperparameter K, the performances of all the clustering methods including expert judgment clustering, are compared. For each clustering method, mortality rates are estimated using the resulting groups of pathologies. The goodness of fit of each mortality model is then assessed by the actual-to-expected ratios as in Section 2.2, and the Bayesian information criterion. The fitted mortality laws of the best model according to the metrics used for the performance comparison, are then analysed in Section 5.3.

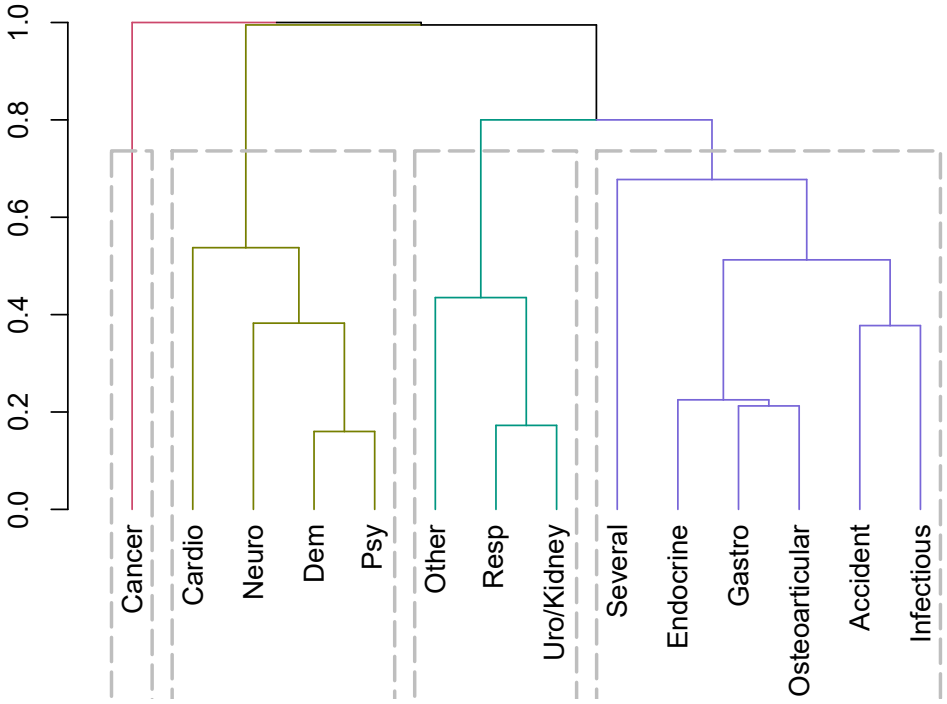

Figure 5: Dendrogram obtained with the complete linkage method from **hclust()** ($K = 4$, Poisson GLM, Formula $= x * Gender + I(x^2) * Gender + t * Gender + I(t^2) * Gender + I(t^3) * Gender$")

5.1 Choice of the number of clusters

The generalized K-means algorithm requires to set the number of final clusters K in advance. This choice therefore has an impact on the final groups and on the estimation of mortality for each pathology. Several methods can be used to select the optimal hyperparameter K. As in Zhang and Lin (2021), our optimization relies on the minimization of the Bayesian information criterion (BIC) proposed by Schwarz (1978), which is given by

$$BIC = -2\log(L) + \log(N)\Lambda, \tag{9}$$

where:

- L denotes the associated maximum likelihood of the model,
- N denotes the number of individuals in the database,
- Λ denotes the number of parameters of the model.

Since the same GLM formula is used for all clusters, Λ is proportional to the number of clusters K in our case. Increasing the number of final clusters leads to an increase in the number of parameters, resulting in an increase in the log-likelihood due to greater flexibility. However, adding complexity can lead to overfitting. The goal of generalized information criterions as the BIC is to find a trade-off between the simplicity and the goodness of fit of the model by adding a penalty that increases with the number of estimated parameters Λ.

Figure 6 shows the evolution of the BIC with K for the two aggregation methods based on generalized K-means. As we can see, allowing for two mortality tables instead of a single common mortality table for all pathologies leads to a significant decrease in the BIC. After decreasing with K, the BIC starts to increase. BIC reaches a minimum for $K = 3$ for the two distance update formulas. Considering only this statistic would lead us to choose $K = 3$. However, the difference in the BIC between $K = 3$ and $K = 4$ is very small. Moreover, for comparison purposes, we prefer to have the same number of final clusters K for all methods, including the GLM tree approach. Therefore, $K = 4$ in the remainder of the paper.

5.2 Comparison of the methods: Goodness of fit

To assess the goodness of fit of each model fitted with the resulting clusters of each clustering method, we first analyse the ratios of actual over expected counts of deaths by gender and pathology. We then compare the Bayesian information criterion computed for each clustering method. Since the GLM tree yields the exact same final groups as the generalized K-means algorithms combined with the complete linkage method for the distance update, only three models are compared:

1. the GLM tree and generalized K-means with complete linkage method,

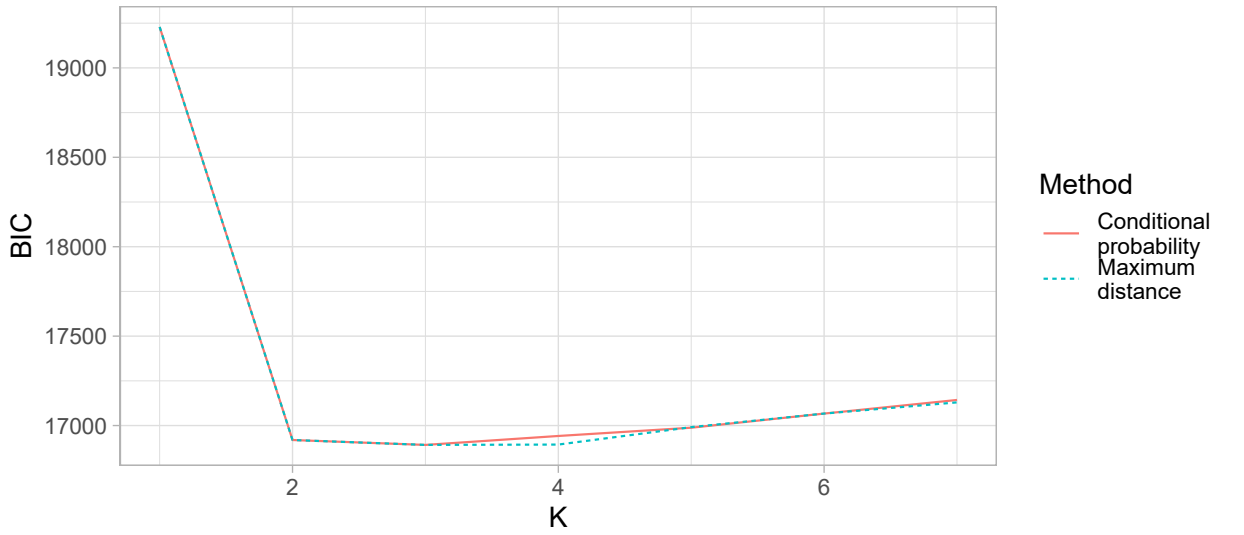

Figure 6: Variation in the Bayesian information criterion with the number of final clusters K

2. generalized K-means combined with the conditional probability method to update the distances, and
3. the expert judgement clustering.

5.2.1 Comparison of the ratios of actual over expected counts of deaths

After estimating mortality rates by using the groups of pathologies resulting from each clustering methods, the performance of each method is assessed by comparing the expected counts of deaths in the portfolio to the observed counts. The ratios of actual over expected counts of deaths by pathology and gender are plotted in Figure 7 for each clustering method, and can be compared to those obtained in the model that does not account for pathology plotted in Figure 1. A model has good predictive performance if the observed number of deaths is close to the expected number of deaths, that is, if the actual-to-expected ratio is close to 1 for each pathology.

Figure 7 shows that the 3 models accounting for pathology, regardless of the clustering method, seem to improve the predictive performance. Let D^p denote a Poisson random variable representing the total number of deaths for pathology p in the database. We want to test the hypothesis

$$\mathcal{H}_0 : D^p \sim Poisson(E^p), \text{ where } E^p \text{ denotes the total expected count of deaths}$$

for each pathology p, with a significance level $\alpha = 5\%$.

Given that only one observation of D^p is available for each pathology p, the well-known chi-squared test, commonly used for Poisson distribution testing, is not suitable. Therefore, another statistical test based on the ratios of actual over expected counts of deaths is constructed to decide whether to

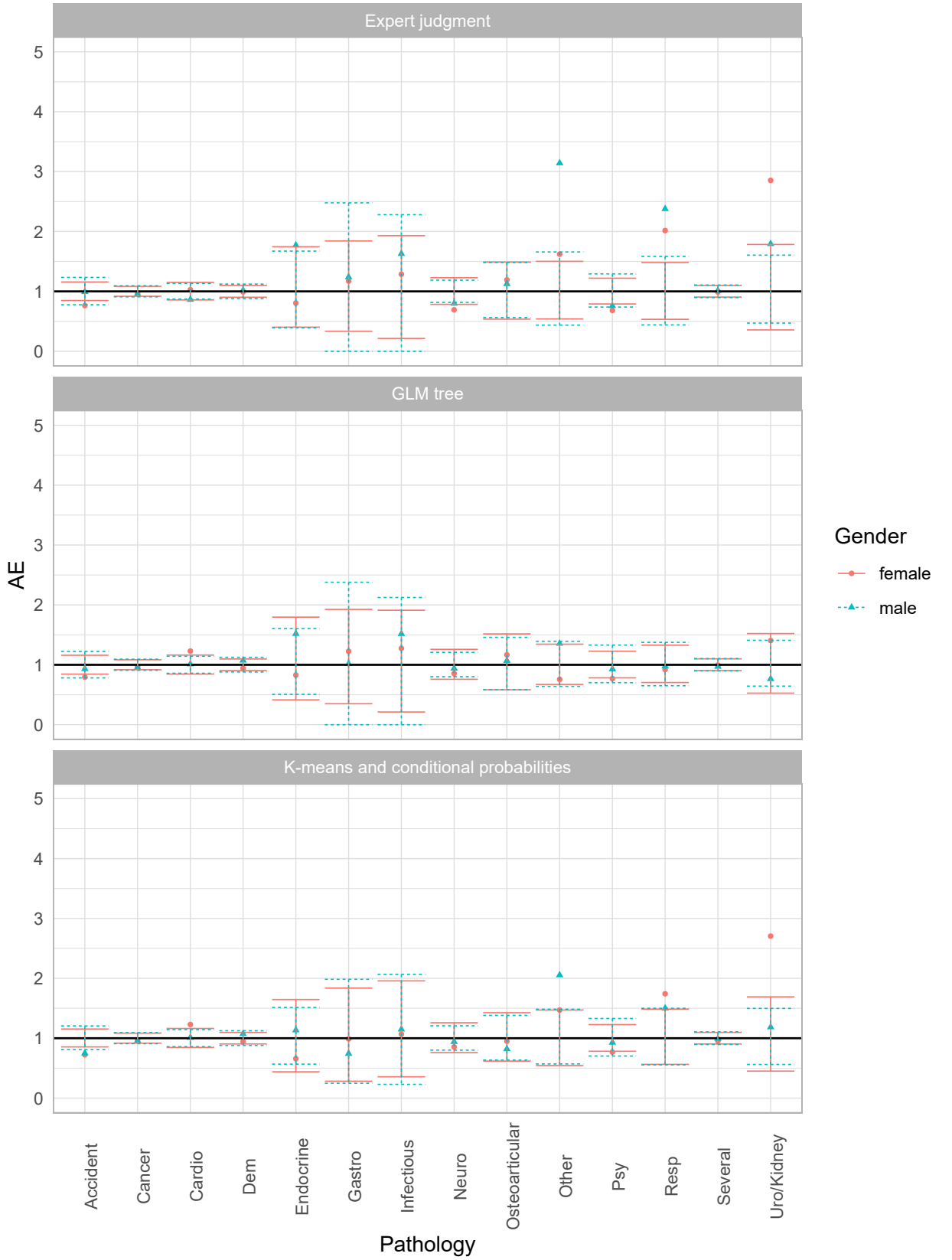

Figure 7: Comparison of the ratios of actual over expected counts of deaths for all clustering methods

reject the null hypothesis $\mathcal{H}_0$ for each pathology.

Let $q_\gamma(E)$ denote the quantile of a Poisson distribution of mean E. $q_\gamma(E)$ is the smallest integer d such that $\mathbb{P}(D \leqslant d) \geqslant \gamma$.

Under $\mathcal{H}_0$,

$$\mathbb{P}\left(\frac{q_{\alpha/2}(E^p)}{E^p} \leqslant \frac{D^p}{E^p} \leqslant \frac{q_{1-\alpha/2}(E^p)}{E^p}\right) \geqslant 1-\alpha,$$

for each pathology p.

A detailed proof of this inequality is given in Appendix A.

Figure 7 shows the bounds of this interval for each pathology p. $\mathcal{H}_0$ is rejected for pathology p, with the signifance level $\alpha = 5\%$, if the ratio of actual over expected counts of deaths is not included in the interval.

Considering the 14 pathologies and two genders, a total of 28 hypotheses were tested for each model. A summary of the statistical hypothesis tests is given in Table 3. For each method, this table summarises the number of rejected hypotheses. Hypothesis $\mathcal{H}_0$ is rejected in 64.3% of the combinations of gender and pathology for the model not accounting for the pathology information. Moreover, the test rejects the hypothesis for highly represented pathologies such as cancer and dementia. The model leading to the fewest rejected hypotheses is the one using the GLM tree clustering method. Only 10.7% of the tests lead to a rejection of $\mathcal{H}_0$. Moreover, $\mathcal{H}_0$ is rejected for females for accident, cardiovascular diseases and psychiatric diseases, representing 6.8%, 5.6% and 2% of the observed claims, respectively. The two other clustering methods also lead to rejection of $\mathcal{H}_0$ for females for accidents and cardiovascular and psychiatric diseases, as well as other combinations of gender and pathology.

Method	**Rejected**	**Not rejected**	**Percent rejected (%)**
Without pathology	18	10	64.3
GLM tree	3	25	**10.7**
K-means combined with conditional probability	9	19	32.1
Expert judgment	12	16	42.9

Table 3: Summary of the statistical hypothesis testing

Therefore, based on these performance metrics, the GLM tree clustering approach seems to be the most suitable method for clustering pathologies in the context of modelling the mortality of disabled policyholders.

5.2.2 Comparison of the Bayesian information criterion

A common metric for comparing models is the Bayesian information criterion (BIC), which was previously used in this paper to choose the optimal number of clusters in Section 5.1. The BIC was computed for each clustering method, as well as for the model not accounting for pathology in the estimation of mortality. The results are summarised in Table 4. The lower BIC is, the better the fitting performance of the model. All models accounting for pathology through clusters have lower BIC values than the model fitting a single mortality law, regardless of the pathology.

Based on this performance metric, the best model is the one using the clusters resulting from the GLM tree method presented in Section 4.1 or from the generalized K-means with the complete linkage method presented in Section 4.2.

Method	**BIC**
Without pathology	19,228.78
GLM tree	**16,893.62**
K-means combined with conditional probability	16,941.71
Expert judgment	17,052.63

Table 4: Comparison of the Bayesian information criterion before and after clustering for each method

Therefore, the model using clusters obtained with the GLM tree method and the generalized K-means combined with the complete linkage method is considered the best model in the remainder of the paper.

5.3 Fitted mortality laws resulting from the best model

The fitted mortality surfaces resulting from the best model are plotted in Figure 8. We recall that although the clusters are identical for males and females, mortality differs with gender. With 4 clusters, we therefore have 8 fitted mortality surfaces depending on attained age and duration, organised as follows: each row represents a cluster, and each column represents a gender. As we can see, the force of mortality associated with cancer is especially high for short durations compared to that associated with other pathologies. Two years after the loss of autonomy, the excess mortality of disabled policyholders with cancer seems to decrease. Dementia and cardiovascular, psychiatric and neurological diseases seem to be the pathologies associated with the lowest mortality after the loss of autonomy for both genders. Disabled policyholders affected by respiratory, urological and kidney diseases are more likely to experience mortality than are those who are disabled because of accidents; endocrine, gastrointestinal, infectious and osteoarticular diseases; or several diseases. The mortality rates resulting from the other clustering approaches are plotted in Appendix B.

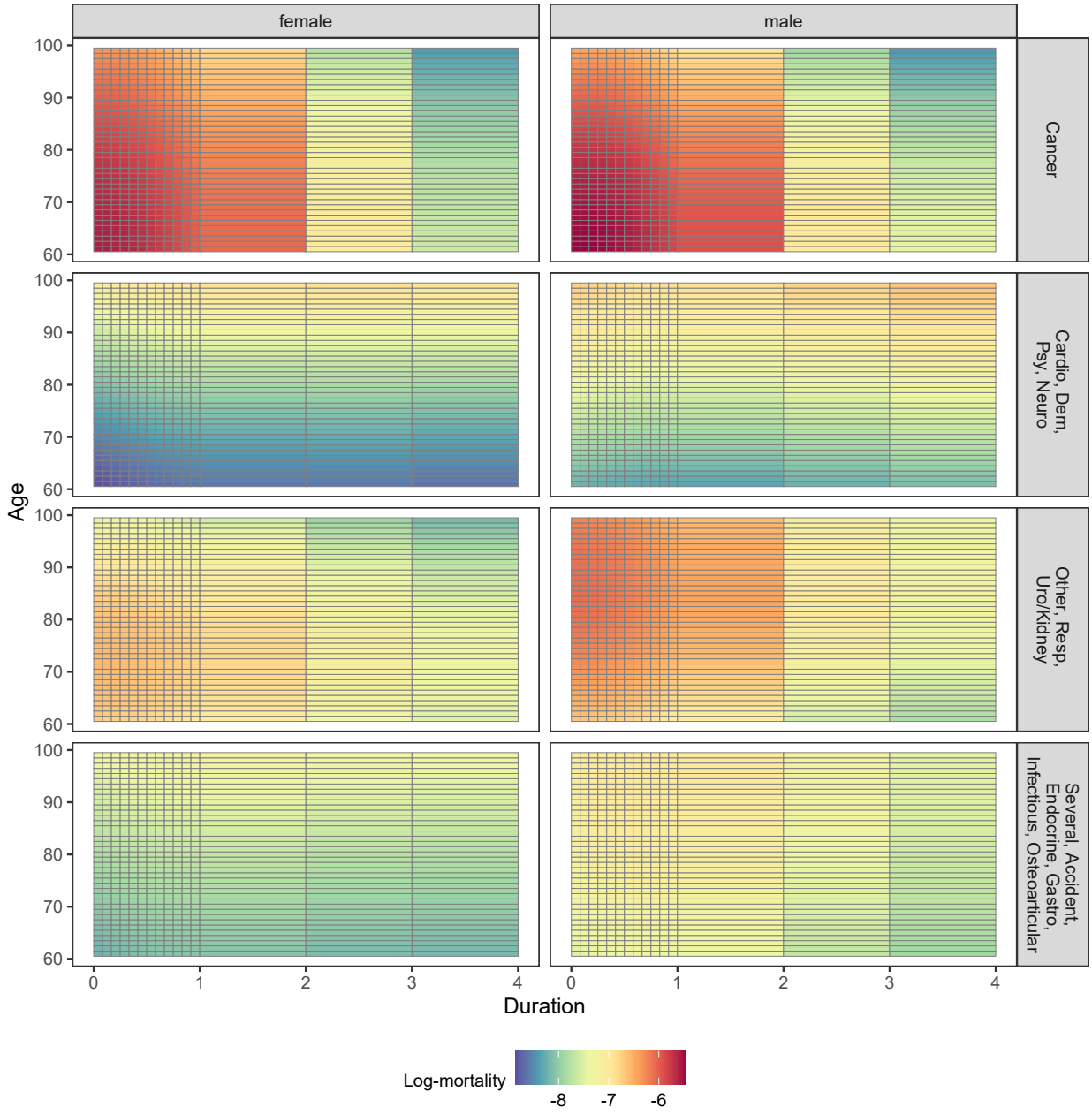

Figure 8: The fitted mortality rates associated to each cluster resulting from the GLM tree or generalized K-means combined with the complete linkage method for updating similarity between clusters

6 Actuarial application

Using the best model chosen in Section 5.2, we can derive some actuarial consequences of accounting for pathology when estimating the mortality of disabled policyholders. We start by estimating the impact of the pathology information on reserving. Then, by accounting for the heterogeneity of mortality between pathologies by deriving 4 different mortality tables associated with the 4 clusters, we can estimate the impact of a change in the incidence rates of a given pathology p on the global mortality in LTC.

6.1 Reserving

Let us consider a LTC insurance product that pays a monthly annuity of $R = 1000€$ in the event of loss of autonomy. The loss of autonomy is assumed to be permanent. Therefore, the only cause of ending annuity payments is the death of the disabled policyholder. Let $i = 3\%$ denote the discount rate. Let $\mu^g_{x,t}$ denote the mortality intensity of a disabled insured individual at age x, knowing his or her gender g and the time since the loss of autonomy t. Let $\mu^{g,p}_{x,t}$ denote the mortality intensity knowing pathology p in addition to the other covariates x, t and g. The actuarial valuation of a lifetime annuity paid to a newly disabled policyholder of age x and gender g is given by the following formula:

$$\Pi^g_x = \sum_{t=1}^{+\infty} \frac{1}{(1+i)^{\frac{t}{12}}} \exp\left(-\frac{1}{12}\sum_{k=0}^{t-1} \mu^g_{x+\frac{k}{12},\frac{k}{12}}\right) R. \tag{10}$$

Knowing the pathology p responsible for his or her loss of autonomy, the valuation of the lifetime annuity paid to the newly disabled policyholder of age x and gender g is given by the following formula:

$$\Pi^{g,p}_x = \sum_{t=1}^{+\infty} \frac{1}{(1+i)^{\frac{t}{12}}} \exp\left(-\frac{1}{12}\sum_{k=0}^{t-1} \mu^{g,p}_{x+\frac{k}{12},\frac{k}{12}}\right) R. \tag{11}$$

Ignoring information about pathology when estimating the mortality of disabled policyholders, an insurer would consider same valuations of the lifetime annuities of all newly disabled policyholders of same age and gender, regardless of the pathologies that caused their loss of autonomy. Let us consider four women who lost their autonomy at age $x = 75$ for distinct causes. The four considered causes of disability are cancer, dementia, respiratory disease and osteoarticular disease. According to the resulting clusters of the GLM tree method, these four pathologies have different mortality rates.

Without accounting for the pathology, the valuations of the lifetime annuities for these four disabled policyholders, computed with Equation 11, are equal to 35,121.68 €.

Accounting for the pathology, the valuations of the annuities of these four disabled policyholders given the pathology are shown in Table 5.

Pathology	$\mathbf{\Pi_x^{g,p}}$	$\mathbf{\frac{\Pi_x^{g,p} - \Pi_x^g}{\Pi_x^g}}$
Cancer	12,009.41€	−65.8%
Dementia	47,130.45€	+34.2%
Resp	23,473.14€	−33.1%
Osteoarticular	40,609.34€	+15.6%

Table 5: Valuation of a lifetime annuity of a disabled woman knowing the pathology

Table 5 shows that the reserves required for newly disabled policyholders of the same age at loss of autonomy strongly depend on the pathology causing their disability. An insurer not accounting for the pathology when estimating reserves would strongly underestimate the resources needed to meet the costs of the loss of autonomy of an insured policyholder with dementia or osteoarticular diseases, as well as any pathology belonging to the same cluster (i.e., accidents; cardiovascular, neurological, endocrine, gastrointestinal, infectious and psychiatric diseases; and several diseases). In contrast, cancer is an aggressive disease that leads to high mortality, especially in the first year following the loss of autonomy, as shown in Figure 8. Therefore, an insurer ignoring pathology information would strongly overestimate the resources needed to meet the costs of the annuity of a disabled policyholder for whom cancer is identified as the cause of loss of autonomy. Considering the pathology, the reserves needed for a policyholder disabled because of cancer are 65.8% lower than the reserves estimated without the information of the pathology.

6.2 Application of a shock

Assuming a unique mortality table for all disabled insured individuals can be sufficient for an insurer if the distribution of the pathologies in the portfolio of disabled policyholders remains the same. The unique mortality table of the disabled individuals can be seen as a weighted average of the mortality of each pathology. The weights are given by the distribution of each pathology at each age and duration. However, due to the heterogeneity of mortality among the different pathologies, a change in the distribution of the pathologies has an impact on the resulting global mortality in LTC. Estimating a unique mortality table regardless of the pathology does not enable an insurer to estimate the impact of a change in the incidence rates associated with a specific pathology. A reduction or increase in the mortality of a specific pathology would also change the distribution of the pathologies in the portfolio, resulting in a modification of the global mortality in LTC.

In this section, we focus on the impact of a reduction or increase in the incidence of a certain pathology. This may occur in the event of preventive actions proposed by the insurer to reduce the incidence of certain pathologies.

Figure 9 shows the impact of an increase or decrease of 10% in the incidence of cancer at each age on mortality during the second year. Reducing the incidence of cancer implies a decrease in the prevalence of this pathology among the disabled policyholders. Since these insured individuals have

a much greater mortality than other disabled policyholders, this change in the incidence implies a decrease in the global mortality in LTC observed in the portfolio. In contrast, increasing the incidence of cancer leads to an increase in the mortality intensity estimated for the second year of disability. As the difference in mortality between clusters is more pronounced at a young age than at greater ages, a variation in the incidence of cancer has a greater impact on global mortality in LTC, all pathologies combined.

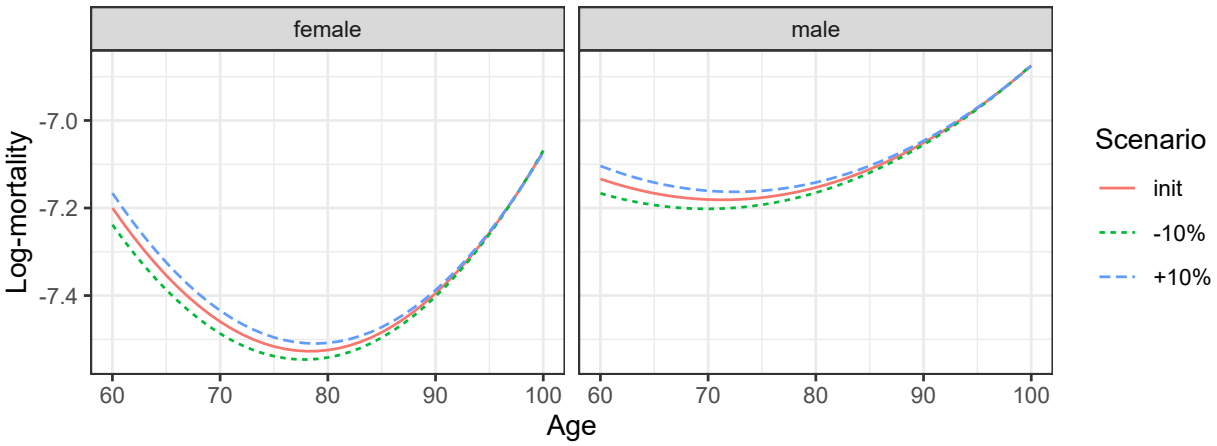

Figure 9: Shocks applied to the incidence of cancer for females: impact on the global mortality in LTC during the second year following the loss of autonomy ($\mu^f_{x,1}$)

The impact of a change in the incidence of dementia at each age on mortality during the second year is plotted in Figure 10. Unlike cancer, dementia is one of the pathologies associated with the highest life expectancy following the loss of autonomy. Therefore, increasing the incidence of dementia increases its prevalence among disabled policyholders, leading to a decrease in the global mortality in LTC.

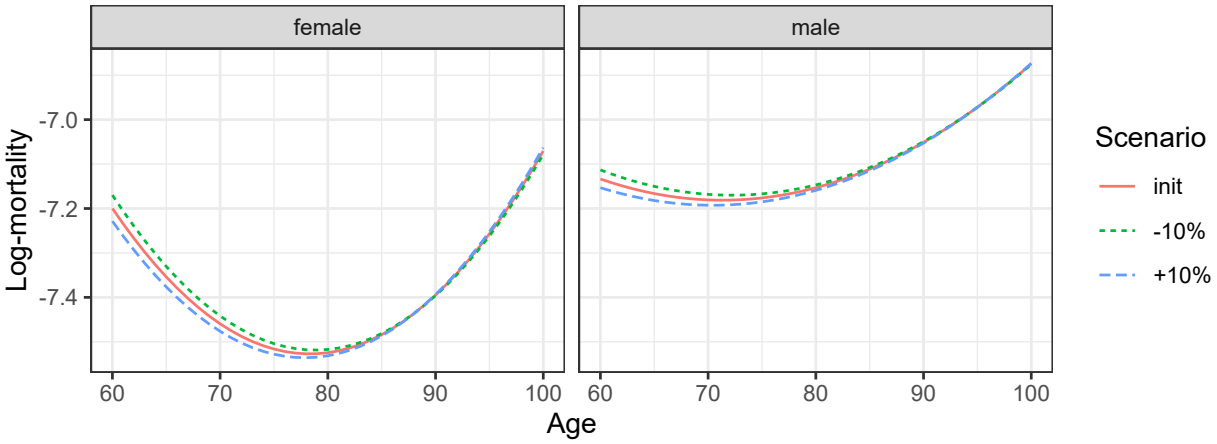

Figure 10: Shocks applied to the incidence of dementia for females: impact on the global mortality in LTC during the second year following the loss of autonomy ($\mu^f_{x,1}$)

The smile shape of the mortality curve observed in Figure 9 and Figure 10 is often observed when estimating the mortality of disabled policyholders for a fixed duration t, especially for low durations. This phenomenon has already been observed in French LTC portfolios, as in Le Bastard et al. (2023). This is mostly due to the evolution of the distribution of pathologies with age among disabled policyholders. The resulting mortality in LTC during the second year of disability, if the loss of autonomy due to cancer was not covered by the insurer, is plotted in Figure 11. The smile shape disappears by setting the incidence of cancer to 0, showing that this specific pattern comes from the prevalence of cancer among the disabled policyholders at young ages.

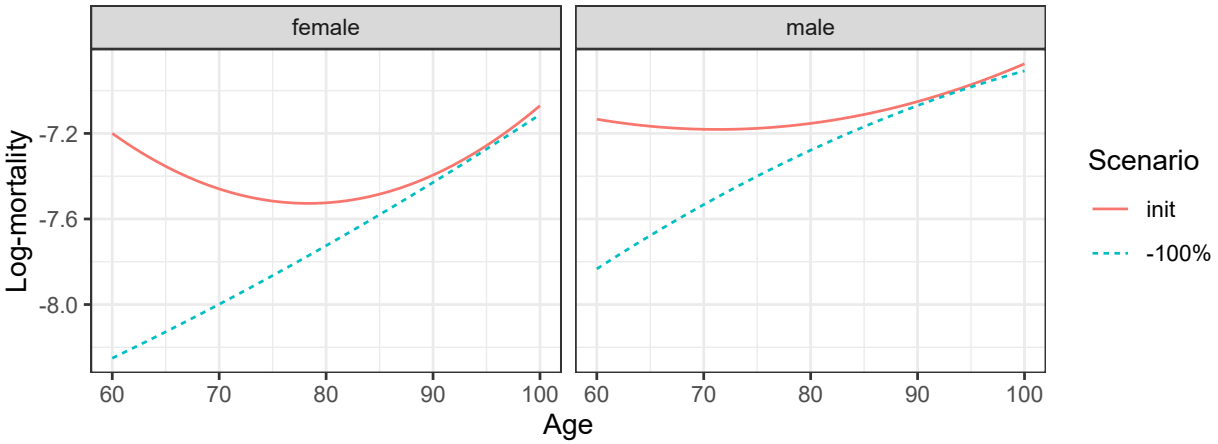

Figure 11: Mortality in LTC during the second year following the loss of autonomy if cancer is excluded from the possible causes of disability

7 Discussion

In the context of studying long-term care insurance products, multiple pathologies can be the cause of an individual's loss of autonomy. Since LTC insurance datasets containing pathology information are rare, literature about the impact of pathology on the mortality of disabled policyholders is still scarce. In this paper, we rely on data from a health fund. After demonstrating the heterogeneity of the force of mortality induced by the plurality of causes of disability, this paper studies two clustering methods to determine which pathologies lead to similar mortality rates after the loss of autonomy. In fact, although accounting for pathology when estimating the mortality of disabled policyholders seems important for capturing the variance, the lack of data does not enable us to independently estimate specific mortality tables depending on gender, attained age and duration of the claim for each pathology. It is therefore necessary to construct groups of homogeneous pathologies. For the sake of simplicity and to reduce the complexity of the models, groups of pathologies are assumed to be identical for both males and females. Despite having the same clusters, the force of mortality is still estimated separately for each gender to account for the heterogeneity between males and females.

Both methods presented in this paper rely on generalized linear models. For each gender and cluster, we assume a quadratic effect of the attained age and a cubic effect of the duration of the claim. The first method uses GLM trees implemented in the **partykit** package in the statistical software **R**. The second method developed in this paper is a generalized K-means approach, resulting in the construction of a similarity measure between pairs of pathologies. Optimal clusters are subsequently constructed using bottom-up hierarchical algorithms. Starting with one cluster for each pathology, each step leads to the merging of the two closest clusters based on a distance measure. The distance measures between each pair of clusters are then updated after each step. Two formulas for updating the distances are tested in this paper. The first relies on the theory of conditional probabilities. The second one, called the complete linkage method, is implemented with the **stats** package in **R**.

As of today, because of the lack of literature on clustering pathologies for long-term care insurance products, most insurers rely on expert judgment to group them. The three clustering methods proposed in this paper, as well as the expert judgment approach, are compared in terms of the goodness of fit of the resulting mortality models. To do this, actual and estimated counts of deaths are compared for each pathology for each clustering approach. The performances of the resulting models are also compared using the Bayesian information criterion. We show that the three clustering methods proposed in this paper yield better fitting performance than does the expert judgment approach. Indeed, experts perform reasoning based on their preconceived view regarding the distribution of the severity of the pathologies in each subgroup of pathologies. An interesting example is that all types of cancer do not have the same severity. Therefore, the clustering is biased by their preconceived view. Using clustering methods based on the portfolio's experience enables the creation of homogeneous groups of pathologies with similar mortality rates, while avoiding the potentially biased expert judgment.

Finally, using the best model based on the two previous performance metrics, we show through actuarial applications the benefits for an insurer of accounting for pathology when estimating the mortality of disabled insureds.

Tables containing the coefficients of the Poisson GLMs fitted using the resulting groups from each clustering method can be found in Appendix C. The p-values presented in the fifth column show the significance of the coefficients related to the variable indicating the cluster, as well as the coefficients associated with interactions of variables involving the cluster. This information again highlights the importance of accounting for pathology as a covariate for predicting mortality.

In the context of modelling LTC products, P-splines are used to fit mortality in LTC in several papers such as Le Bastard et al. (2023), which enables to consider more complex structure of the surface that describes mortality. As mortality rates vary greatly between different pathologies, the distribution of pathologies among disabled insureds in the portfolio significantly impacts the global mortality in LTC. Although the shape of the mortality surface in LTC is complex, part of this complexity comes from the variation in the distribution of pathologies across age and duration. Therefore, a simpler impact of age and duration given the pathology is assumed in this paper, by

considering an additive impact of age and duration. A cross effect of age and duration has been considered, but did not improve the performance of the models. Access to data containing more observations would allow us to consider using P-splines.

We remind the reader that this work relies on data coming from a foreign market having a specific definition of the LTC insurance. The resulting groups of pathologies may not be transposable to other countries for two main reasons: the heterogeneity of the populations and potential differences in the definition of the loss of autonomy. However, most current data in France do not include detailed information on the pathology. While our work provides an initial insight on groups of similar pathologies, future research should study the sensitivity of the constructed groups of pathologies with respect to the country.

Future research should also consider different groups of pathologies for males and females. Another idea would be to consider different groupings depending on attained age and duration. To this end, we suggest using the survival trees described in Bou-Hamad et al. (2011). In this method, all covariates can be partitioned recursively to create groups of homogeneous observations.

A Appendix A: Proof of the statistical hypothesis test for the actual-to-expected ratios

We want to test the hypothesis

$$\mathcal{H}_0 : D \sim Poisson(E) \text{ where } E \text{ denote the expected count of deaths,}$$

with a significance level α.

Let $\gamma \in [0; 1]$. Let $q_\gamma(E)$ denotes the quantile of a Poisson distribution with parameter E. $q_\gamma(E)$ is the smallest integer d such that

$$\mathbb{P}\left(D \leqslant d\right) \geqslant \gamma.$$

Therefore we have

$$\mathbb{P}\left(D < q_{\tilde{\gamma}}(E)\right) < \tilde{\gamma}.$$

Let γ and $\tilde{\gamma}$ denote two probability values; then,

$$\begin{aligned}\mathbb{P}\left(q_{\tilde{\gamma}}(E) \leqslant D \leqslant q_\gamma(E)\right) &= \mathbb{P}\left(D \leqslant q_\gamma(E)\right) - \mathbb{P}\left(D < q_{\tilde{\gamma}}(E)\right) \\ &\geqslant \gamma - \tilde{\gamma}.\end{aligned}$$

Let $\gamma = 1 - \alpha/2$ and $\tilde{\gamma} = \alpha/2$; then,

$$\mathbb{P}\left(q_{\alpha/2}(E) \leqslant D \leqslant q_{1-\alpha/2}(E)\right) \geqslant 1 - \alpha,$$

and

$$\mathbb{P}\left(\frac{q_{\alpha/2}(E)}{E} \leqslant D \leqslant \frac{q_{1-\alpha/2}(E)}{E}\right) \geqslant 1 - \alpha. \tag{12}$$

Therefore, the null hypothesis $\mathcal{H}_0$ is rejected if 1 is not included in the interval $\left[\frac{q_{\alpha/2}(E)}{E}; \frac{q_{1-\alpha/2}(E)}{E}\right]$, with a significance level α.

B Appendix B: Plots of mortality rates resulting from the nonselected clustering approaches

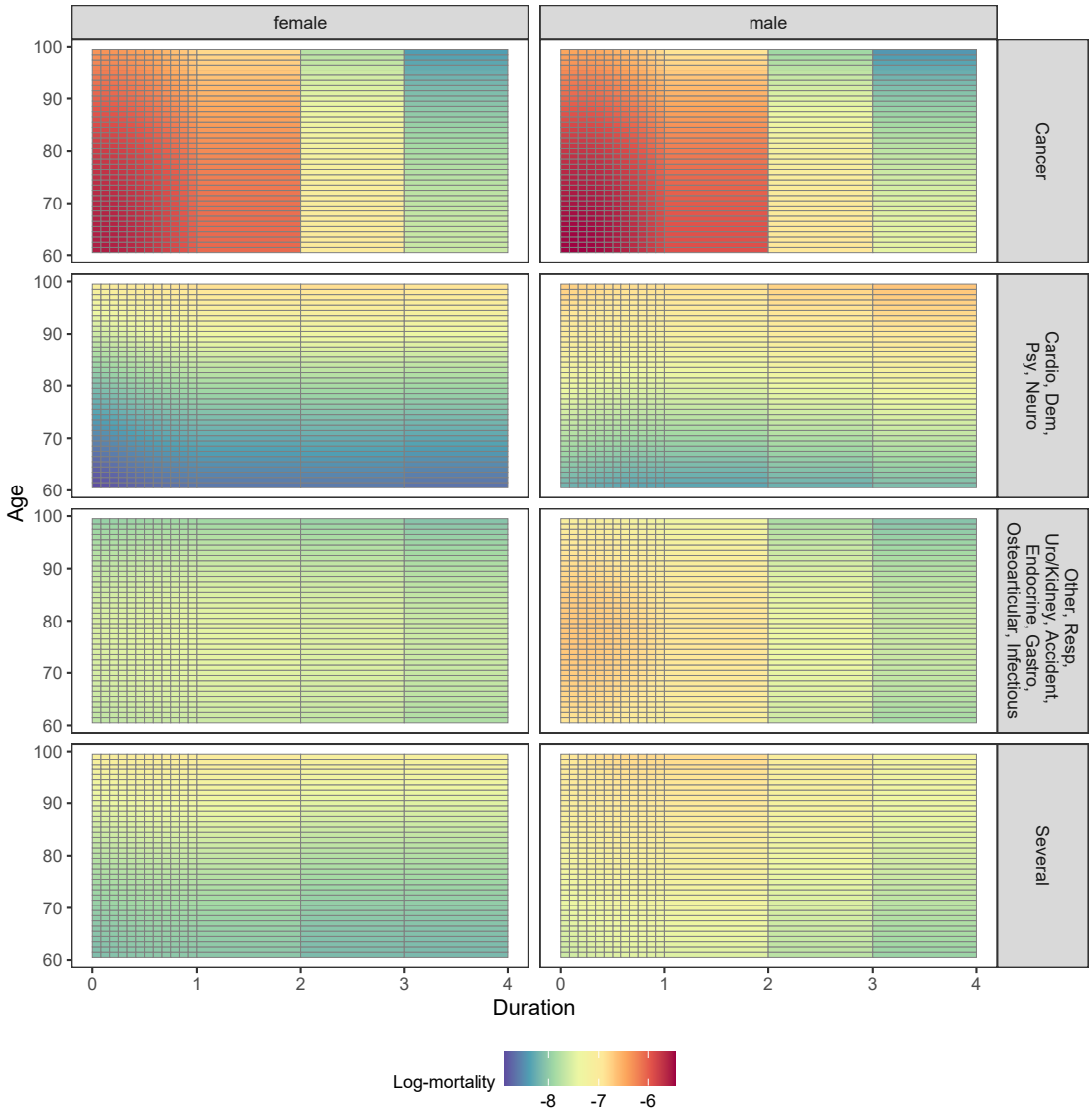

Figure 12: The fitted mortality rates associated to each cluster resulting from the generalized K-means algorithm with distance measure using conditional probabilities

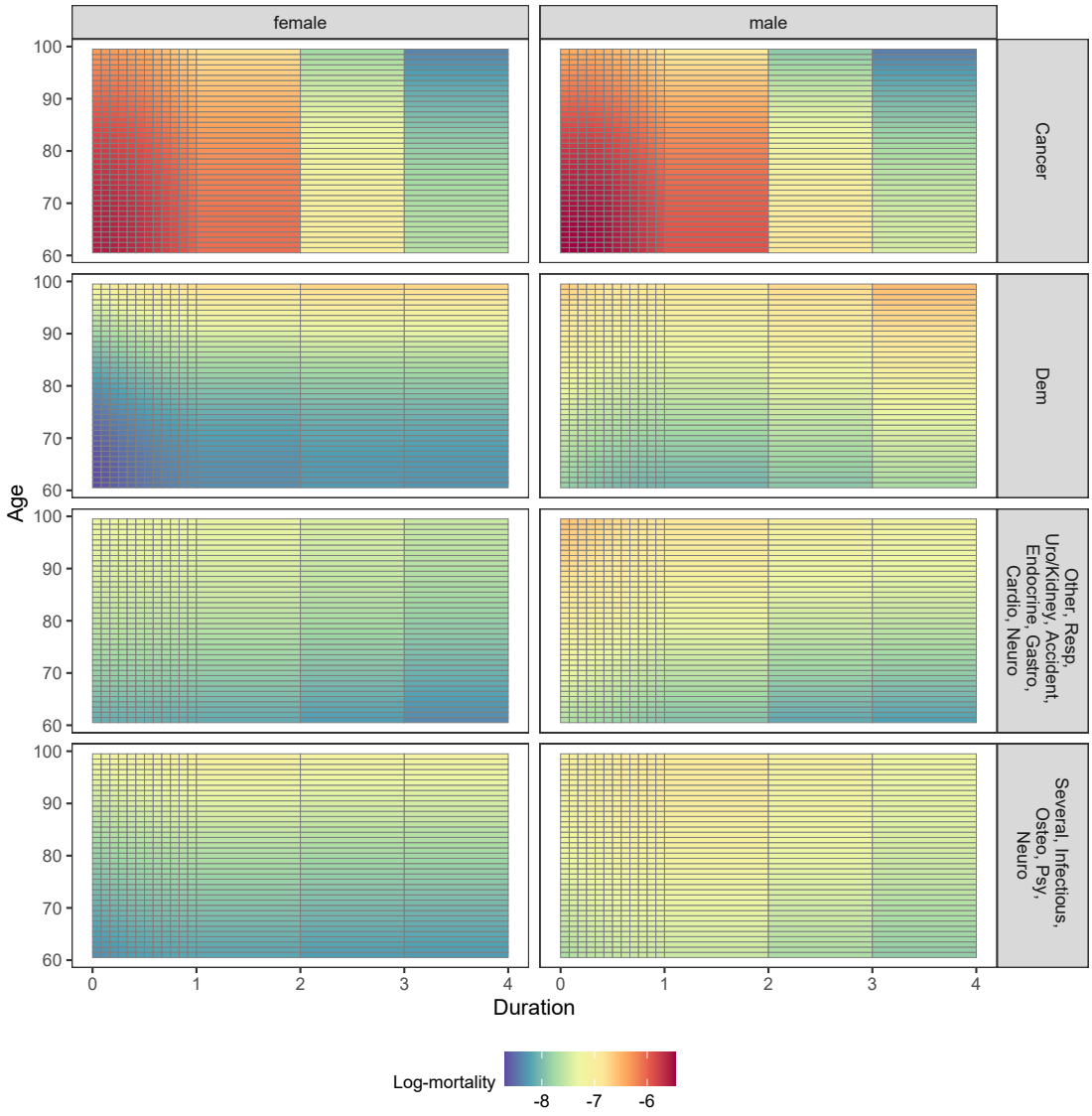

Figure 13: The fitted mortality rates associated to each cluster resulting from the expert judgment approach

C Appendix C: Coefficients of the Poisson GLM using pathology groups from each clustering methods

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-9.60	0.96	-10.05	0.00	***
x	-0.01	0.02	-0.35	0.73	
Gendermale	-2.54	1.67	-1.52	0.13	
cluster2	0.48	1.25	0.38	0.70	
cluster3	2.40	1.03	2.32	0.02	*
cluster4	-2.72	1.39	-1.95	0.05	·
I(x^2)	0.00	0.00	2.11	0.03	*
t	0.43	0.23	1.90	0.06	·
I(t^2)	-0.21	0.13	-1.63	0.10	
I(t^3)	0.03	0.02	1.48	0.14	
x:Gendermale	0.09	0.04	2.08	0.04	*
x:cluster2	0.02	0.03	0.60	0.55	
x:cluster3	0.06	0.03	2.30	0.02	*
x:cluster4	0.17	0.04	4.18	0.00	***
Gendermale:cluster2	2.15	1.96	1.10	0.27	
Gendermale:cluster3	2.09	1.76	1.19	0.23	
Gendermale:cluster4	2.23	2.16	1.03	0.30	
Gendermale:I(x^2)	0.00	0.00	-2.18	0.03	*
cluster2:I(x^2)	0.00	0.00	-1.29	0.20	
cluster3:I(x^2)	0.00	0.00	-4.24	0.00	***
cluster4:I(x^2)	0.00	0.00	-4.78	0.00	***
Gendermale:t	-0.74	0.33	-2.21	0.03	*
cluster2:t	-0.11	0.31	-0.34	0.73	
cluster3:t	-0.43	0.31	-1.36	0.17	
cluster4:t	-0.52	0.65	-0.80	0.42	
Gendermale:I(t^2)	0.43	0.20	2.18	0.03	*
cluster2:I(t^2)	-0.01	0.18	-0.05	0.96	
cluster3:I(t^2)	-0.41	0.22	-1.85	0.06	·
cluster4:I(t^2)	0.03	0.44	0.07	0.94	
Gendermale:I(t^3)	-0.07	0.03	-2.06	0.04	*
cluster2:I(t^3)	0.01	0.03	0.27	0.79	
cluster3:I(t^3)	0.10	0.04	2.51	0.01	*
cluster4:I(t^3)	0.01	0.08	0.12	0.91	
x:Gendermale:cluster2	-0.05	0.05	-1.05	0.29	
x:Gendermale:cluster3	-0.07	0.05	-1.53	0.13	
x:Gendermale:cluster4	-0.10	0.06	-1.60	0.11	
Gendermale:cluster2:I(x^2)	0.00	0.00	0.95	0.34	
Gendermale:cluster3:I(x^2)	0.00	0.00	1.45	0.15	
Gendermale:cluster4:I(x^2)	0.00	0.00	1.87	0.06	·
Gendermale:cluster2:t	0.82	0.46	1.78	0.08	·
Gendermale:cluster3:t	0.91	0.49	1.86	0.06	·
Gendermale:cluster4:t	0.76	0.89	0.85	0.39	
Gendermale:cluster2:I(t^2)	-0.61	0.28	-2.19	0.03	*
Gendermale:cluster3:I(t^2)	-0.60	0.37	-1.62	0.11	
Gendermale:cluster4:I(t^2)	-0.63	0.60	-1.05	0.29	
Gendermale:cluster2:I(t^3)	0.10	0.05	2.19	0.03	*
Gendermale:cluster3:I(t^3)	0.11	0.07	1.55	0.12	
Gendermale:cluster4:I(t^3)	0.12	0.10	1.12	0.26	

Table 6: Coefficients of the Poisson GLM using clusters from the GLM trees

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-12.21	0.82	-14.86	0.00	***
x	0.12	0.02	5.29	0.00	***
Gendermale	0.31	1.02	0.30	0.76	
cluster2	5.00	0.91	5.49	0.00	***
cluster3	2.61	1.26	2.07	0.04	*
cluster4	4.65	1.27	3.66	0.00	***
I(x^2)	0.00	0.00	-4.79	0.00	***
t	0.18	0.35	0.50	0.61	
I(t^2)	-0.11	0.22	-0.49	0.62	
I(t^3)	0.01	0.04	0.34	0.73	
x:Gendermale	0.01	0.03	0.43	0.67	
x:cluster2	-0.07	0.03	-2.57	0.01	*
x:cluster3	-0.13	0.03	-3.89	0.00	***
x:cluster4	-0.15	0.03	-4.58	0.00	***
Gendermale:cluster2	-0.76	1.16	-0.66	0.51	
Gendermale:cluster3	-2.85	1.96	-1.45	0.15	
Gendermale:cluster4	-2.10	1.87	-1.12	0.26	
Gendermale:I(x^2)	0.00	0.00	-0.32	0.75	
cluster2:I(x^2)	0.00	0.00	1.75	0.08	·
cluster3:I(x^2)	0.00	0.00	4.93	0.00	***
cluster4:I(x^2)	0.00	0.00	5.06	0.00	***
Gendermale:t	-0.12	0.49	-0.25	0.80	
cluster2:t	-0.17	0.41	-0.42	0.68	
cluster3:t	0.26	0.42	0.61	0.54	
cluster4:t	0.08	0.43	0.18	0.86	
Gendermale:I(t^2)	-0.26	0.31	-0.85	0.39	
cluster2:I(t^2)	-0.52	0.28	-1.81	0.07	·
cluster3:I(t^2)	-0.10	0.26	-0.41	0.68	
cluster4:I(t^2)	-0.11	0.26	-0.42	0.67	
Gendermale:I(t^3)	0.07	0.05	1.33	0.18	
cluster2:I(t^3)	0.12	0.05	2.33	0.02	*
cluster3:I(t^3)	0.02	0.04	0.42	0.67	
cluster4:I(t^3)	0.03	0.04	0.69	0.49	
x:Gendermale:cluster2	0.01	0.03	0.17	0.87	
x:Gendermale:cluster3	0.08	0.05	1.47	0.14	
x:Gendermale:cluster4	0.05	0.05	1.07	0.28	
Gendermale:cluster2:I(x^2)	0.00	0.00	-0.31	0.75	
Gendermale:cluster3:I(x^2)	0.00	0.00	-1.51	0.13	
Gendermale:cluster4:I(x^2)	0.00	0.00	-1.16	0.25	
Gendermale:cluster2:t	0.30	0.61	0.49	0.62	
Gendermale:cluster3:t	-0.61	0.59	-1.03	0.30	
Gendermale:cluster4:t	0.29	0.62	0.48	0.63	
Gendermale:cluster2:I(t^2)	0.09	0.44	0.21	0.83	
Gendermale:cluster3:I(t^2)	0.69	0.37	1.89	0.06	·
Gendermale:cluster4:I(t^2)	0.12	0.39	0.31	0.76	
Gendermale:cluster2:I(t^3)	-0.03	0.08	-0.36	0.72	
Gendermale:cluster3:I(t^3)	-0.14	0.06	-2.21	0.03	*
Gendermale:cluster4:I(t^3)	-0.05	0.06	-0.77	0.44	

Table 7: Coefficients of the Poisson GLM using clusters from the K-means algorithm with the distance measure using conditional probabilities

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.20	0.40	-18.23	0.00	***
x	0.06	0.01	4.36	0.00	***
Gendermale	-0.45	0.54	-0.84	0.40	
cluster2	-4.07	0.79	-5.15	0.00	***
cluster3	2.35	3.11	0.75	0.45	
cluster4	-1.84	0.95	-1.93	0.05	·
I(x^2)	0.00	0.00	-4.61	0.00	***
t	0.00	0.22	0.02	0.98	
I(t^2)	-0.62	0.18	-3.47	0.00	***
I(t^3)	0.13	0.03	3.87	0.00	***
x:Gendermale	0.02	0.02	1.06	0.29	
x:cluster2	0.02	0.02	0.67	0.50	
x:cluster3	-0.18	0.08	-2.37	0.02	*
x:cluster4	-0.05	0.03	-2.12	0.03	*
Gendermale:cluster2	0.21	1.06	0.19	0.85	
Gendermale:cluster3	-3.81	4.24	-0.90	0.37	
Gendermale:cluster4	-0.54	1.28	-0.43	0.67	
Gendermale:I(x^2)	0.00	0.00	-1.07	0.28	
cluster2:I(x^2)	0.00	0.00	0.97	0.33	
cluster3:I(x^2)	0.00	0.00	3.18	0.00	**
cluster4:I(x^2)	0.00	0.00	3.61	0.00	***
Gendermale:t	0.17	0.36	0.49	0.63	
cluster2:t	0.04	0.34	0.11	0.92	
cluster3:t	0.59	0.38	1.56	0.12	
cluster4:t	0.33	0.32	1.02	0.31	
Gendermale:I(t^2)	-0.17	0.31	-0.55	0.59	
cluster2:I(t^2)	0.55	0.24	2.28	0.02	*
cluster3:I(t^2)	0.35	0.25	1.40	0.16	
cluster4:I(t^2)	0.40	0.22	1.78	0.08	·
Gendermale:I(t^3)	0.04	0.06	0.67	0.50	
cluster2:I(t^3)	-0.12	0.04	-2.75	0.01	**
cluster3:I(t^3)	-0.09	0.04	-2.07	0.04	*
cluster4:I(t^3)	-0.09	0.04	-2.25	0.02	*
x:Gendermale:cluster2	0.00	0.03	-0.12	0.90	
x:Gendermale:cluster3	0.12	0.11	1.17	0.24	
x:Gendermale:cluster4	0.03	0.03	0.94	0.35	
Gendermale:cluster2:I(x^2)	0.00	0.00	0.44	0.66	
Gendermale:cluster3:I(x^2)	0.00	0.00	-1.22	0.22	
Gendermale:cluster4:I(x^2)	0.00	0.00	-1.01	0.31	
Gendermale:cluster2:t	-0.54	0.51	-1.05	0.29	
Gendermale:cluster3:t	-1.28	0.58	-2.19	0.03	*
Gendermale:cluster4:t	-0.03	0.50	-0.06	0.95	
Gendermale:cluster2:I(t^2)	0.27	0.39	0.70	0.48	
Gendermale:cluster3:I(t^2)	0.78	0.42	1.87	0.06	·
Gendermale:cluster4:I(t^2)	0.04	0.38	0.10	0.92	
Gendermale:cluster2:I(t^3)	-0.05	0.07	-0.64	0.52	
Gendermale:cluster3:I(t^3)	-0.13	0.08	-1.73	0.08	·
Gendermale:cluster4:I(t^3)	-0.02	0.07	-0.30	0.76	

Table 8: Coefficients of the Poisson GLM using clusters from expert judgment

References

Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Achim Zeileis, T. H. and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

Barndorff, N. (1978). Information and exponential families; in statistical theory. Technical report.

Biessy, G. (2016). A semi-Markov model with pathologies for Long-Term Care Insurance. preprint.

Biessy, G. (2017). Continuous-time semi-markov inference of biometric laws associated with a long-term care insurance portfolio. *ASTIN Bulletin: The Journal of the IAA 47*(2), 527–561.

Biessy, G. (2019). Smoothing of multidimensional biometric laws in a long-term care insurance portfolio.

Bou-Hamad, I., D. Larocque, and H. Ben-Ameur (2011). A review of survival trees.

Czado, C. and F. Rudolph (2002). Application of survival analysis methods to long-term care insurance. *Insurance: Mathematics and Economics 31*(3), 395–413.

Debón, A., L. Chaves, S. Haberman, and F. Villa (2017). Characterization of between-group inequality of longevity in european union countries. *Insurance: Mathematics and Economics 75*, 151–165.

Fuino, M. and J. Wagner (2018). Long-term care models and dependence probability tables by acuity level: New empirical evidence from switzerland. *Insurance: Mathematics and Economics 81*, 51–70.

Guibert, Q. and F. Planchet (2018). Non-parametric inference of transition probabilities based on Aalen–Johansen integral estimators for acyclic multi-state models: application to LTC insurance. *Insurance: Mathematics and Economics 82*, 21–36.

Hoem, J. M. (1972). Inhomogeneous semi-Markov processes, select actuarial tables, and duration-dependence in demography. In T. Greville (Ed.), *Population Dynamics*, pp. 251–296. Academic Press.

Hothorn, T. and A. Zeileis (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research 16*(1), 3905–3909.

Hunt, A. and D. Blake (2021). On the structure and classification of mortality models. *North American Actuarial Journal 25*(sup1), S215–S234.

Janssen, J. (1966). Application des processus semi-markoviens à un probléme d'invalidité. *Bulletin de l'Association Royale des Actuaries Belges 63*, 35–52.

Le Bastard, L., S. Loisel, and A. W. Shao (2023). Combining experience data of several Long-Term Care Insurance products with different disability definitions. working paper or preprint.

Léger, A.-E. and S. Mazzuco (2021). What can we learn from the functional clustering of mortality data? An application to the Human Mortality Database. *European Journal of Population 37*.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Morone, F. (2022). Clustering matrices through optimal permutations. *Journal of Physics: Complexity 3*(3), 035007.

Myers, R. H. and D. C. Montgomery (1997). A tutorial on generalized linear models. *Journal of Quality Technology 29*(3), 274–291.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Pitacco, E. (2014). Health insurance. *Basic Actuarial Models, Cham, Switzerland: Springer Verlag*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

RDocumentation (2023). hclust: Hierarchical clustering. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461 – 464.

Soetewey, A., C. Legrand, M. Denuit, and G. Silversmit (2022, 04). Semi-Markov modeling for cancer insurance. *European Actuarial Journal 12*.

Xuanyuan Shihao and Shiang Xuanyuan (2023). Application of Markov model in long-term care insurance. *Highlights in Science, Engineering and Technology 47*, 9–15.

Zhang, T. and G. Lin (2021). Generalized k-means in GLMs with applications to the outbreak of COVID-19 in the United States. *Computational Statistics Data Analysis 159*, 107217.