View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 2, 2020 (New Series), pp 333–351

Finite Mixture Modelling of Survival Data: With Applications to Pensioner Lifetime

Jyrki Möttönen[1], Janne Salonen[2], Tapio Nummi[3] and Timothy E. O'Brien[4]
[1]*Department of Mathematics and Statistics, University of Helsinki, P.O.Box 68, FI-00014 University of Helsinki, Finland*
[2]*Finnish Centre for Pensions, 00065 Eläketurvakeskus, Helsinki, Finland*
[3]*Faculty of Information Technology and Communication Sciences, Tampere University, Finland*
[4]*Loyola University Chicago, Department of Mathematics and Statistics, and Institute of Environmental Sustainability, USA*

Received: 01 April 2020; Revised: 09 September 2020; Accepted: 11 September 2020

Abstract

One of the most popular methods for modelling survival analysis data is the ubiquitous and time-honored proportional hazards model of Cox (1972). Popularity notwithstanding, in several cases, however, the proportional hazards assumption is found to be violated. Thus, important model extensions have been developed in the intervening years. One such extension is the so-called Frailty model (see for example Collett, 2015), which is based on the utilization of random effects. Another extension arises from the application of finite mixtures in the context of time-to-event data. Although finite mixture modelling tools are used in many fields of science, they have been less well-developed and used in survival analysis. Thus, a key aim of this article is to provide an interesting application of mixture modeling in survival analysis and to discuss aspects arising in its application. In this paper, we apply these techniques using real data from the research register of the Finnish Centre for Pensions; using pension insurance mortality data, we use the basic Cox proportional hazards model by incorporating finite mixture modelling techniques. Additional comparisons with frailty models are also provided.

Key words: Censoring; Cox model; Disability pension; EM algorithm; Finite mixtures; Mixture models; Mortality.

1. Introduction

Survival analysis (SA) techniques include a set of methods for analyzing time until the occurrence of a pre-specified event of interest such as mortality. In SA, subjects are usually followed over a pre-specified period of time. As such, an event can be, for example, death, the occurrence of a disease or the end of working life. SA can also be used to handle incomplete information. This is called censoring. Observations are censored when the information

Correponding Author: Tapio Nummi
Email: tapio.nummi@tuni.fi

about their survival time is incomplete. A commonly encountered form of censoring is right censoring, which means that the event of interest did not occur during the chosen follow-up period. When studying lifetime data, SA techniques are often considered superior to normal linear regression. SA is applied in many fields of science, including economics (Heckman and Singer, 1985; van den Berg, 2001), medicine and health (Machin *et al.*, 2006; Collett, 2015), amongst other fields.

One of the most popular method or model in SA is the Cox proportional hazards (CPH) model (Cox, 1972). In important cases, however, the required proportional hazards assumption is violated and this has led to key CPH model extensions. For example, in some cases the survival times among individuals are not independent. This situation may arise if individuals falling into specific groups (such as hospitals or clinics) tend to follow similar survival times. These types of model extensions are often addressed in Frailty models (see Collett, 2015). The basic idea in these models is to apply mixed modelling (with random effects) in a SA context. In this type of modelling, the source of correlation is assumed to be known. Frailty modelling includes two main approaches (see *e.g.*, Wienke, 2010). First, it is assumed that the event times include clusters, defined as shared frailties (random effects). Second, frailty can also be defined on an individual level, in which case the corresponding model is called the univariate frailty model. A comprehensive treatment of random effects in survival modelling is given in Ha *et al.* (2017) for example.

In another type of modelling, the observed survival times may be correlated, but the source of correlation cannot be directly measured. Basically, this kind of situation can be addressed using the theory and methods of finite mixtures (*e.g.*, McLachlan and Peel, 2000). One technique used in mixture longitudinal data analysis is trajectory analysis (*e.g.*, Nagin, 2005; Nagin and Odgers, 2010a), where the focus is on the analysis of a sequence of measurements. This technique has been widely applied in the social sciences (Nagin and Odgers, 2010b; Nummi *et al.*, 2017), but applications to survival data have not been that common. Thus, although finite mixture modelling tools are used in many fields of science (*e.g.*, Böhning *et al.*, 2007), they are not that much utilized with CPH model especially. Some related topics are covered in textbooks, like Ng *et al.* (2019) and McLachlan and Peel (2000), but there are surprisingly few published articles of the topic. Obviously, there are some practical obstacles in the combination of these two techniques and we try to figure out what these are with interesting heterogeneous pension insurance survival data that is used for testing and illustration.

In this paper, we employ real data from the research register of the Finnish Centre for Pensions. In the Finnish context, the causes of disability correlate with mortality as highlighted in Polvinen *et al.* (2015) and Sewdas *et al.* (2020). Using these data and focusing on mortality, we test and illustrate several important and practical modelling scenarios. In the first scenario, we apply the basic Cox proportional hazards model. In the second scenario, we apply mixture modelling to survival times, and then we use the identified mixture components as a risk factor in the basic Cox model. Our third scenario includes applying mixture modelling within Cox model. The final scenario involves using a frailty model with individual-level frailties. Our results are summarized in the final section as well as recommendations for future research.

2. Theoretical Background

2.1. Finite mixture analysis

A finite mixture of censored T distribution regression models is defined by letting G denote the number of mixture components (*i.e.*, groups) in the mixture model and Z_i denote the latent class (random) variable which indicates the component (group or sub-population) to which the ith observation (individual) belongs. The conditional density of the outcome variable Y_i, $i = 1, \ldots, n$, given $Z_i = j$, is given by

$$Y_i|\{z_i = j\} \sim T(\beta_0 + \mu_j + \mathbf{x}_i^T\beta, \sigma_j^2, \nu), \quad j = 1, \ldots, G, \tag{1}$$

where $(\beta_0 + \mu_j)$ is the intercept of jth group, $\mathbf{x}_i$ is a p-dimensional predictor vector, β is a p-dimensional parameter vector, σ_j is the scale parameter, ν is the degrees of freedom and $T(\mu, \sigma, \nu)$ is the generalized T distribution with location parameter μ, scale parameter σ and ν degrees of freedom. Note that if $Y \sim T(\mu, \sigma, \nu)$, then the distribution of $(Y - \mu)/\sigma$ is standardized Student's T distribution with ν degrees of freedom. Clearly, the mean of the density within the jth group is $\beta_0 + \mu_j + \mathbf{x}_i^T\beta$.

In the case of right censoring, the observed ith outcome can be defined as follows:

$$Y_i^* = \begin{cases} c_i, & \text{if } \rho_i = 1 \ (i.e.\ y_i > c_i) \\ y_i, & \text{if } \rho_i = 0. \end{cases} \tag{2}$$

Here,

$$\rho_i = \begin{cases} 1, & \text{if the } i\text{th observation is right-censored,} \\ 0, & \text{if the } i\text{th observation is not censored.} \end{cases}$$

Suppose that $P(Z_i = j) = p_j$ and Z_i is independent of predictor variables $\mathbf{x}_i$. The maximum likelihood estimate is then obtained by maximizing the log-likelihood function

$$l(\theta|y_1^*, \ldots, y_n^*, \rho_1, \ldots, \rho_n) = \sum_{i=1}^{n} \log\{p_j[f_{ij}(y_i^*)]^{1-\rho_i}[1 - F_{ij}(y_i^*)]^{\rho_i}\} \tag{3}$$

where f_{ij} and F_{ij} are the probability density function and cumulative distribution function of $T(\beta_0 + \mu_j + \mathbf{x}_i^T\beta, \sigma_j^2, \nu)$ distribution. Correspondingly, $\theta = (\theta_1^T, \ldots, \theta_G^T)$, where $\theta_j = (p_j, \beta_j^T, \sigma_j^2, \nu)^T$. In the R-package CensMixReg, the maximum likelihood estimate of θ is found using an EM-type algorithm (Dempster *et al.*, 1977). Note that it is also possible to fit a finite mixture of censored normal distribution regression models using CensMixReg. However, in our empirical experience and the context of our application, using the T distribution gave more consistent results than the normal distribution.

The posterior probability of the ith individual belonging to the jth mixture component is estimated by

$$p_{ij} = \frac{p_j[f_{ij}(y_i^*)]^{1-\rho_i}[1 - F_{ij}(y_i^*)]^{\rho_i}}{\sum_{h=1}^{G} p_h[f_{ih}(y_i^*)]^{1-\rho_i}[1 - F_{ih}(y_i^*)]^{\rho_i}}, \quad i = 1, \ldots n, \ j = 1, \ldots, G. \tag{4}$$

The individual i is assigned to the group having the highest posterior probability estimate.

Likelihood-based inference can be very difficult in mixture models and this is also the case when mixtures and Cox models are combined. For example, when testing hypotheses $H_0 : \ G = G_0$ against $H_A : G = G_0 + 1$, the usual likelihood ratio testing should not be applied because the clusters may not be nested. Also, other point identification and boundary problems may appear. However, the usual information criteria and some other statistical measures may still be applied.

2.2. The proportional hazards (PH) and frailty model

2.2.1. The hazard function

Let Y be the time to an event. The hazard function can be defined as follows:

$$h(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y < y + \Delta y | Y \geq y)}{\Delta y}. \tag{5}$$

In this expression, $P(y \leq Y < y + \Delta y | Y \geq y)$ is the conditional probability that the event occurs in a short time-interval, given that the event has not occurred before time y. Note that $h(y)$ is not the probability that the event occurs at time y or before time y. We can interpret the hazard function $h(y)$ as an instantaneous rate of occurrence of an event (*e.g.*, death). We can also approximate the conditional probability with $h(y)\Delta y$, where Δy is a small positive real number.

2.2.2. The Cox PH model with finite mixtures

The Cox proportional hazards model can be defined in terms of the hazard function in the following manner:

$$h_i(y) = \lambda_0(y) \exp(\beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \tag{6}$$

where $h_i(y)$ is the hazard of individual i at time y, $x_{i1}, \ldots, x_{ik}$ are k covariates of the individual i, $\beta_1, \ldots, \beta_k$ are the model regression coefficients and $\lambda_0(y)$ is the baseline hazard function.

The above model is called the proportional hazards model because the hazard ratio relating individual i to individual j,

$$\frac{h_i(y)}{h_j(y)} = \exp\{\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk})\}, \tag{7}$$

does not depend on time nor on the base hazard function $\lambda_0(y)$. In his groundbreaking paper, Cox (1972) showed that the regression coefficients can be estimated using partial likelihood methods without knowing the form of the base hazard.

In the mixture modelling context the population density function of time to event has the finite mixture form

$$f(y) = \sum_{j=1}^{G} \pi_j f_j(y), \tag{8}$$

where $f_1(y), \ldots, f_G(y)$ are the densities of the mixture components and $\pi_1, \ldots, \pi_G$ are the mixing proportions which add up to one. Then the survival function of the time to the event has the mixture form

$$S(y) = \sum_{j=1}^{G} \pi_j S_j(y), \tag{9}$$

where $S_1(y), \ldots, S_G(y)$ are the survival functions of the mixture components. However, it has been shown that the hazard function of the time to the event does not have a mixture form under CPH (McLachlan and McGiffin, 1994; Ng *et al.*, 2019). Another possibility is to use a mixture specification of the hazard function

$$h(y) = \sum_{j=1}^{G} \pi_j h_j(y), \tag{10}$$

where $h_1(y), \ldots, h_G(y)$ are the hazard functions of the mixture components, but in that case the survival function of the time to the event does not have a mixture form (McLachlan and McGiffin, 1994; Ng *et al.*, 2019).

Eng and Hanlon (2014) have proposed a method where the mixture components are estimated using EM-algorithm. If the observation y of the jth mixture component follows PH model, it has the density

$$f_j(y, \delta|\mathbf{x}) = [\lambda_{0j}(y) \exp(\mathbf{x}^T \beta_j)]^{\delta} \exp[-H_{0j}(y) \exp(\mathbf{x}^T \beta_j)], \tag{11}$$

where δ is the censoring indicator ($\delta = 1$, if the survival time is observed), $\mathbf{x}$ is the covariate vector, y is the survival time, $\lambda_{0j}(y)$ is the baseline hazard, $H_{0j}(y)$ is the cumulative hazard and β_j is the regression coefficient vector. The density of the complete data can then be written as

$$f(\mathbf{y}, \boldsymbol{\Delta}|\mathbf{x}, \mathbf{U}) = \prod_{i=1}^{n} \sum_{j=1}^{G} [\pi_j f_j(y_i, \delta_i|\mathbf{x}_i)]^{u_{ij}}, \tag{12}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ are the survival times, $\boldsymbol{\Delta} = (\delta_1, \ldots, \delta_n)$ are the censoring indicators ($\delta_i = 1$, if the ith survival time is observed), $x_i = (x_{i1}, \ldots, x_{ik})^T$ is the covariate vector, $\mathbf{U} = (U_1, \ldots, U_n)$ are the latent mixture components ($U_i = j$, if the ith observation belongs to the jth component), $P(U_i = j) = \pi_j$ and $u_{ij} = 1_{\{U_i = j\}}$. Eng and Hanlon (2014) maximized the mixture likelihood using EM-algorithm and called the method *Cox-assisted clustering* (*CAC*).

2.2.3. The shared frailty PH model

The shared frailty model is defined as follows:

$$h_{ij}(y) = h_0(y) \exp(\mathbf{x}_{ij}^T \beta + w_j) = h_0(y) u_j \exp(\mathbf{x}_{ij}^T \beta),\ j = 1, \ldots, G,\ i = 1, \ldots, n_j, \tag{13}$$

where w_j, $j = 1, \ldots, G$, are i.i.d. random effects distributed as $N(0, \sigma^2)$, $h_{ij}(y)$ is the conditional hazard of individual i from the jth component (conditional on w_j), $\mathbf{x}_{ij}$ is the

vector of covariates, β are regression coefficients and $h_0(y)$ is the baseline hazard function. The $u_j = e^{w_j}$ term is the frailty of the jth component, where $j = 1, \ldots, G$. Note that now $\log(u_j)$ is normally distributed. It follows that the hazard ratio is

$$\frac{h_{i_2 j_2}(y)}{h_{i_1 j_1}(y)} = \frac{u_{j_2}}{u_{j_1}} \exp((\mathbf{x}_{i_2 j_2} - \mathbf{x}_{i_1 j_1})^T \beta), \quad j_1 \neq j_2. \tag{14}$$

From this expression, it is seen that the hazard ratio depends on the frailties. We can easily see that if the individual i_2 from the j_2th component and the i_1 from the j_1th component have identical covariate profiles (*i.e.* $\mathbf{x}_{i_2 j_2} = \mathbf{x}_{i_1 j_1}$), then the hazard ratio simplifies to

$$\frac{h_{i_2 j_2}(y)}{h_{i_1 j_1}(y)} = \frac{u_{j_2}}{u_{j_1}}. \tag{15}$$

Note that if the number of components is the same as the number of individuals (*i.e.* $n = \Sigma_j n_j = G$), we get the following individual frailty model or random effects frailty model:

$$h_i(y) = h_0(y) \exp(\mathbf{x}_i^T \beta + w_i) = h_0(y) u_i \exp(\mathbf{x}_i^T \beta), \quad i = 1, \ldots, n. \tag{16}$$

3. Pension Insurance Mortality Data

The practical application of this study relates to pensioners and their mortality. The research data was collected from the research register of the Finnish Centre for Pensions. The register is a national databank of the Finnish population, including both the working age population and retirees. The databank contains comprehensive socioeconomic information on the population and statutory pensions.

The base population consisted of the subset of individuals born in 1940 who were still alive in 1995 (*i.e.*, aged 55). The data is cross-sectional from an analytical point of view but entails longitudinal follow-up information on lifetime from 1995 to 2018 (*i.e.*, from ages 55 to 78). Here, we used a 50-per-cent random sample from the selected cohort and this translated into a total of 10,637 individuals. The remaining lifetime of this cohort underlines the difference in mortality between men and women since, according to Official Statistics of Finland (2020), the expected years alive after 2019 for men was 8.9 years and for women 10.9 years.

The research data contains individual-level information on the following variables:

- Lifetime in years (from 55th birthday until 31 Dec. 2018),
- Gender,
- Pension benefit (Disability Pension or Old-age Pension),
- Cause of permanent disability leading to disability pension (8 classes),
- Employer before retirement (Private Sector, Public Sector or Self-employed),
- Censoring (1=alive or 0=deceased on 31 Dec. 2018),

- Highest Education (Basic Education, Secondary Education, Lower University Degree or Higher University Degree), and
- Age at retirement (a continuous variable while the others are classifying variables).

Lifetime is the response variable that we analyzed in this study, using both finite mixture and survival modelling techniques.

Our analysis focuses on mortality. We followed the individuals of the cohort born in 1940 for the 24-year-period from 1995 to 2018 (*i.e.*, ages 55 to 78). This follow-up time translated into a great share of the cohort alive at the end of the study period (over 72%), disregarding possible illnesses. Furthermore, in the Finnish population (as in many other western countries), the probability of permanent disability, which ends working lives for the majority of the seriously ill people, increases rapidly after age 55. However, a vast majority of the cohort survived without permanent illnesses and could retire on an old-age pension at the agreed retirement age of the pension scheme (65 yrs.). From the perspective of SA, the censoring rate in the population was 72.8 per cent.

The actual data analysis was performed using R software (R Core Team, 2019). Specifically, the CensMixReg package (Sanchez *et al.*, 2018) was used to identify mixture groups, and the Survival (Therneau and Lumley, 2019) and Coxme (Therneau, 2020) survival analysis R packages. Furthermore, we have also used the *cac* R function of Eng and Hanlon (2014) to fit the Cox PH model within the mixture model context.

Table 1 shows some basic descriptive statistics within the respective factor classes. The counts indicate a reasonable number of cases in the classes in the sense that the chosen categories are not so fine as to result in sparse data. Specifically, the smallest class, higher university graduates ($n = 134$), is non-censored, and the other classes include more than 300 individuals.

The shares of the deceased (non-censored) indicate that mortality among men and disability pensioners is high, as high share of these groups face death before age 78. The share of non-censored is significantly lower among women (20%), old-age pensioners (19%) and lower for university graduates (20%).

The same classes are reflected in the average lifetimes of the non-censored. When comparing men and women, the lifetime (by age 78) of the non-censored women is 0.9 years longer than that of men. The same difference can be seen when comparing disability pensioners with old-age pensioners. The lifetime of disability pensioners is 3.9 years shorter than that of old-age pensioners. When comparing classes of education, the lifetimes are relatively similar within the three lowest classes: basic education (15.6 yrs.), secondary education (15.4 yrs.), and lowest university education (15.8 yrs.). The lifetime of high university graduates is 1.1 years longer than that of those with a basic education. The classes of employer before retirement indicate significant but small differences between classes.

Descriptive statistics indicate that we can expect significant differences between genders and pension benefits in the statistical analyses given in the following section.

Table 1: Descriptive statistics of cohort born in 1940

	Non-censored			Censored	Total
	Count	%	Average Lifetime, years*,**	Count	Count
Men	1,803	36	15.3 [15.1–15.5]	3,269	5,072
Women	1,087	20	16.2 [15.9–16.4]	4,478	5,565
Old-age pension	1,300	19	17.8 [17.6–18.0]	5,635	6,935
Disability pension	1,590	43	13.9 [13.6–14.1]	2,112	3,702
Basic ed.	1,721	30	15.6 [15.3–15.8]	3,995	5,716
Secondary ed.	676	27	15.4 [15.1–15.8]	1,856	2,532
Low univ.	359	20	15.8 [15.3–16.3]	1,400	1,759
High univ.	134	21	16.7 [16.0–17.5]	496	630
Private Sector	1,490	30	15.3 [15.0–15.5]	3,481	4,971
Public Sector	1,005	24	15.9 [15.6–16.2]	3,247	4,252
Self-employed	395	28	16.1 [15.7–16.6]	1,019	1,414
Total	2,890			7,747	10,637

*90% confidence limits in parentheses.
**Average lifetime from 55th birthday until 31 Dec. 2018.

4. Data Analysis Using Various Modelling Techniques

4.1. The Cox model

The first step of our study was to analyze our data using a Cox proportional hazards model. The response variable was Lifetime and the explanatory variables were Gender, Pension benefit, Age at retirement, Highest education and Employer. The model estimates (Hazard ratios) are given in Table 4 (under the column titled "Cox PH Fit"). The analysis was done both for the whole data set and for the data set with censored cases excluded. The first analysis includes 10,637 individuals, of which 2,890 are non-censored. The results are reasonable and mirror much of what was demonstrated in Table 1. From the descriptive statistics given in Table 1, significant differences in lifetime in two variables, Gender and Pension benefit, are observed. From the Cox PH model analysis (Table 4) we observe that women's mortality rate is significantly lower (*i.e.*, hazard ratio 0.5) compared to men's mortality rates. Also, disability pensioners have a much higher hazard ratio of mortality (2.42) than old-age pensioners. In comparison with individuals in the basic education level, those with a higher university education have a slightly lower mortality rate (the coefficient for

higher university education (0.82) is statistically significant). Employer is not a statistically significant variable in this model.

The second analysis was done for the non-censored data set to make the analysis and goodness-of-fit measures comparable to the results of the Cox PH with ex ante (in the SA, we call the groups as "ex ante groups" since they are constructed using the survival model underlying equation (17) below. The concordance statistics is an established measure of the goodness-of-fit in survival models (see Harrell *et al.*, 1996); for the basic Cox model, the concordance value is 0.61. Subsequent sections show differences between other models and Cox PH model.

4.2. Modelling mixture components

The second step of our analysis was to first study the distribution of the lifetime. The lifetime is largely centered around 23–24 years, indicating the still living individuals. Furthermore, the long left tail indicates increasing mortality with age. The first step of finite mixture modelling was to search for possible mixture components or sub-groups from the outcome. In general, choice of the number of mixture components is a key central question in finite mixture modelling, and it is usually determined via statistical information criteria (*e.g.*, Bayesian Information Criteria, BIC) and with the subjective consideration of the modeller. In our application, the number of groups was based on the sizes of the mixture components, as we did not want possible artefact or too small groups in the further steps of the analysis.

When using the CensMixReg R package (and included functions), which takes account of right-censoring on the outcome, a regression model must be specified. To this end, we defined the following simple regression model:

$$Lifetime = PensionBenefit + RetirementAge + HighestEducation \tag{17}$$

The distribution of the outcome Lifetime is such that the R implementation standard assumption of the normal distribution was switched to the T distribution, which behaved slightly more stable in our analyses. Some experiments with gender as a factor indicated that the model is somewhat sensitive to the underlying regression model (factors), and so we chose to use a simple model, which provided us with a reasonable number of mixture components and which could be estimated with the EM algorithm. The BIC values for $k = 2, 3, 4$ groups were respectively 28757.09, 28799.61 and 28955.36. As is often the case, choices of a larger number of mixture components/groups led to convergence problems of the EM algorithm, and were therefore not considered here. Overall in the BIC analysis the model solution was stable between several model runs.

The four-component solution yielded the group sizes as shown in Table 2. We named the groups based on average lifetime, counted from the data using the group-assignments. The No risk group included those who were alive at the end of the 23-year study period, and thus had censored measurements in the survival modelling analysis.

To further illustrate the above lifetime analysis, we draw the distribution of lifetimes by

Table 2: Mixture group sizes and lifetimes

	Estimated Counts	%	Average Lifetimes (years)
No Risk (NR)	7,747	72.8	> 23.6
Low Risk (LR)	169	1.6	22.7
At Risk (AR)	325	3.1	20.1
High Risk (HR)	2,396	22.5	14.5

mixture component. The boxplot of lifetimes for the mixture components are shown in Figure 1. As demonstrated by Figure 1 and Table 2, the components or sub-groups have greatly different mortalities. The vast majority of the sample (72.8%) are in the No Risk group, and alive at the end of study period. The Low Risk and At Risk groups show a slightly increased mortality with an average lifetimes of 22.7 years and 20.1 years and relatively narrow range of lifetimes. The average lifetime in High Risk group is only 14.5 years and the Figure shows also a wide range of lifetimes. The overall conclusion from Figure 1 is that the distribution between mixture components barely overlap indicating a clear group assignment of the above mixture analysis.

Although the distribution of lifetimes barely differ for men and women within components, the shares or proportions of men and women within the components are different, as there are significantly more men in the High Risk group (see Table 3 below). These additional results give further assurance for the group-based correlation of survival times and we will utilize this information in conjunction with the Cox PH regression model in the following section.

Table 3: Mixture groups by gender

	Men		Women		Total	
	Count	%	Count	%	Count	%
No Risk (NR)	3,269	64.5	4,478	80.5	7,747	72.8
Low Risk (LR)	104	2.1	65	1.2	169	1.6
At Risk (AR)	196	3.9	129	2.3	325	3.1
High Risk (HR)	1,503	29.6	893	16.0	2,396	22.5
Total	5,072		5,565		10,637	

The mixture groups can be further dissected and parsed by analyzing the cause of retirement on a disability pension. The cause of permanent disability gives a clear indication or a proxy of an individual's health, and some causes of disability are more life-threatening than others. For example, in examining Figure 2, we see that neoplasms are often life-threatening in the sense that those presenting with neoplasms are represented to a large degree in the High Risk mortality group. Indeed, approximately 74 per cent of individuals with neoplasms were in the High Risk mortality group. Conversely, individuals with depression or diseases of the musculoskeletal system were somewhat over-represented in the No Risk group. Interestingly, although the cause of disability was not used in mixture analysis, this analysis revealed substantial agreement with these results.

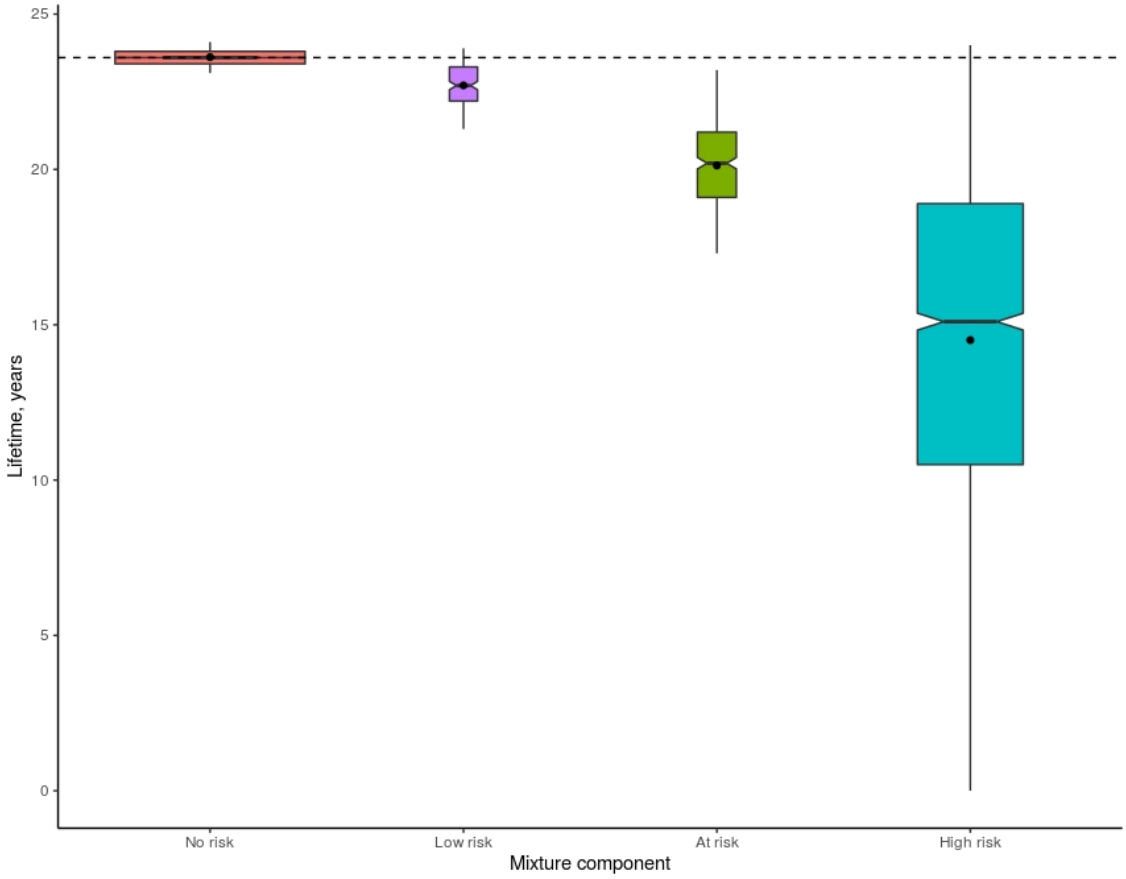

Figure 1: Box plot of lifetime by mixture component. Box widths are proportional to sample sizes. The notches indicate the sample medians and the dots indicate the sample means. The dashed line indicates the sample median of the censored observations.

4.3. The Cox model with mixture components as a factor

The third step in our analysis was to combine the four-class factor obtained in the preceding section with the basic Cox PH model developed in Section 4.1. To achieve this, we used the risk groupings from the information of ex ante defined mixture groups. The hazard ratios are also presented in Table 4 in the column labelled "Cox PH + Ex Ante Fit". For those individuals who died before the end of the 23-year follow-up period, we select one class of the mixture components as a reference group, and in this analysis, the chosen reference group was the Low Risk group. As noted in Table coefficients, the fitted results have changed in some meaningful ways as compared with the Cox PH model fit. For example, now the hazard ratio of disability pensioners is much higher (7.00) compared to old-age pensioners. The mixture-component-based classes are statistically significant in the model. The coefficients are reasonable and confirm that the mixture groups capture interesting sub-populations. The estimates for the At Risk group (3.03), and especially for the High Risk group (24.00), are large compared to the reference (Low Risk) group, as is to be expected.

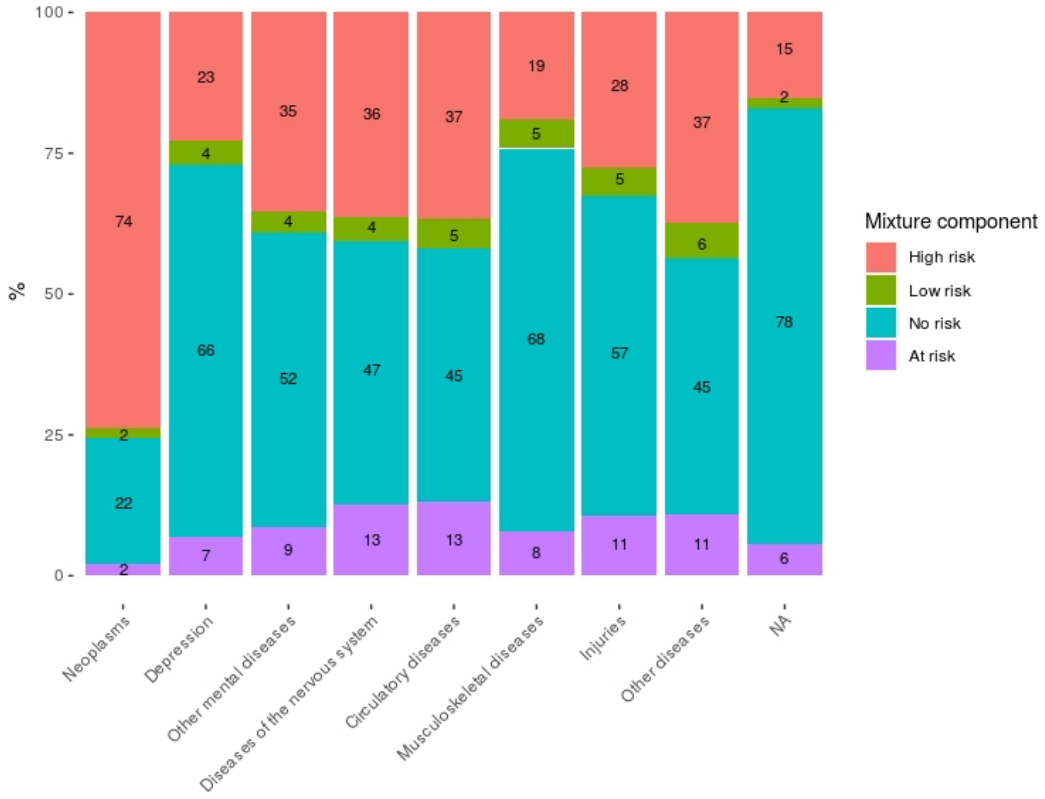

Figure 2: Cause of disability by mixture component, %

In comparing the basic Cox PH model fit with the Cox PH combined with Ex Ante fit (treating groups as a factor), the LR test statistic χ^2 value is 1754.9 on 2 degrees of freedom. This difference is highly significant and this indicates that adding the group component information significantly increases the predictive power of the Cox PH model. Similarly, the concordance statistic (0.73) for the latter model is higher than the value corresponding to the Cox PH model (0.61), which also indicates an improved model fit.

4.4. The frailty model with random effects

The fourth step of our analysis included fitting a frailty model with individual random effects. The results are listed in Table 4 (Column titled "Frailty with Random Effects Fit"). Note that in this analysis we did not use the mixture component factor developed in Section 4.2 and used in Section 4.3. The results can be summarized by noting that the hazard ratios are quite similar to the basic Cox PH model. The coefficient for women (0.47) is slightly smaller than in the Cox PH model (0.50) and significantly smaller than in the Cox model with mixture components as a factor (0.92), and the coefficient for the disability pension (2.67) is slightly higher than in the Cox PH model (2.42). As was the case for the basic Cox PH model, the Frailty model disability pension estimate differs greatly from the Cox PH model with mixture components, which yields the estimated hazard ratio of 7.0.

Table 4: Cox and frailty regression estimates (hazard ratios)

		Cox PH Fit		Cox PH + Ex Ante Fit	Frailty with Random Effects Fit
		Censored included	Censored excluded	Censored excluded	Censored included
Gender:	Women	0.50***	0.91*	0.92***	0.47***
	Men	ref	ref	ref	ref
Pension:	Old-age	ref	ref	ref	ref
	Disability	2.42***	1.50***	7.00***	2.67***
Pension age		0.98***	0.99	0.97***	0.98***
Education:	Basic ed.	ref	ref	ref	ref
	Secondary ed.	0.92	1.05	0.88**	0.91*
	Low univ.	0.77***	0.99	0.78***	0.75***
	High univ.	0.82*	0.95	0.84	0.79*
Employer:	Private sector	ref	ref	ref	ref
	Public sector	1.06	0.97	0.95	1.06
	Self-employed	0.94	0.96	0.94	0.93
Group:	No Risk	N/A	N/A		N/A
	Low Risk	N/A	N/A	ref	N/A
	At Risk	N/A	N/A	3.03***	N/A
	High Risk	N/A	N/A	24.00***	N/A
RE Variance					0.60
Concordance			0.61	0.73	
The number of cases (n)		2890	2890	2890	2890

Significance codes: $***< 0.001$, $**< 0.01$, $*< 0.05$

The basic Cox PH model test against the frailty mode were compared with a LR test. The value of the statistic was 33.978 (1 degrees of freedom), which clearly indicates that the frailty model with individual effects increases the predictive power of the Cox model.

To further develop and illustrate this model, we provide the estimated random coefficients (*i.e.*, individual frailties) in Figure 3 in the Appendix. The estimated values for the No Risk group (mean = -0.16) are all negative highlighting that the individuals in this group survived until the end of the study period (*i.e.*, until age 78). Essentially all of the High Risk (0.48), At Risk (0.24) and Low Risk (0.13) group estimates are positive. Not surprisingly, the range of the coefficients do not differ appreciably between the Low-Risk and At-Risk groups, but the coefficients for the High Risk group are notably higher (mean and median around 0.5) than in the other groups.

4.5. The mixture Cox model

Our last analysis consists of mixture modelling of lifetime using the finite mixture form of the Cox PH model as shown in section 2.2.2. The number of mixture groups is more complicated in our application as far more groups are separated or indicated as compared to the ex ante modelling of lifetime as in section 4.2. The BIC analysis focusing on non-censored cases indicate at least 15 groups with relatively similar survival patterns. The survival times of the groups are similar and near each other. To simplify the results, we selected a three-group solution with relatively equal sizes. Also 4 group solution was tested, but the magnitude of the estimates obtained was no longer at a plausible level.

Table 5 shows the Cox model estimates within the mixture groups. The gender (woman) effect is highly group-specific and estimates also differ compared to basic Cox model (Table 4, Censored excluded). Estimates on disability pension and pension age are in the same range between groups and, similar to basic Cox model. Estimates of level of education differ greatly between groups. Overall the statistical significance of most of the covariates is high, especially in groups two and three. Figure 4 in the Appendix shows the corresponding distribution of lifetime within mixture groups. The figure shows that the lifetimes are relatively similar between groups. In the high mortality group one the average lifetime is 14.1 years and correspondingly 15.8 years and 18.2 years in groups two and three. The model selected here includes three sub-groups (Non-censored only). However, we can see that the groups are different from the ex ante mixture analysis (see Table 2) where groups indicate very different mortalities. The ex ante mixture analysis showed a large High Risk group ($N = 2,396$) with high mortality (average lifetime 14.5 yrs.) and two smaller groups with moderate mortality. In the mixture Cox model the groups are somewhat more similar and not easily comparable with the ex ante groups.

5. Concluding Remarks

Survival analysis (SA) techniques are appealing approaches in many fields of research and application because of their simplicity, their nonparametric nature, and their accommodation of the presence of censoring. The key question regarding the ubiquitous Cox model is whether the resulting model estimates are unbiased. To address such deficiencies, frailty models have been developed to include a random factor for unknown covariates, and the frailties aim to capture effects which are not explicitly included in the model. Furthermore, the merit of the finite mixture modelling approach is that it reveals possible latent classes from a given outcome distribution. As a starting point, as is demonstrated in our work here it is noteworthy and interesting to bring the sub-groups or mixture components into the context of SA. Using these techniques in parallel will continue to provide new insights into the Cox model, especially regarding estimates and dealing with any related biases.

In this study, we analyzed a simple empirical data, including information on lifetimes and some background factors of a Finnish cohort. The focus of the analysis was on modelling lifetime at ages 55 to 78. The basic Cox model revealed differences in mortality with respect to gender, pension benefit and education. The results indicate that women face a smaller likelihood of death compared to men, and disability pensioners face a far greater likelihood of death than old-age pensioners.

Table 5: Mixture Cox model regression estimates by mixture group (hazard ratios)

		Group 1	Group 2	Group 3
Gender:	Women	0.01***	7.91***	14.93***
	Men	ref	ref	ref
Pension:	Old-age	ref	ref	ref
	Disability	4.26***	2.64***	4.63***
Pension age		0.99	0.86***	1.04***
Education:	Basic ed.	ref	ref	ref
	Secondary ed.	0.94	0.02***	24.17***
	Low univ.	3.14***	0.004***	11.35***
	High univ.	0.01***	0.07***	82.79***
Employer:	Private sector	ref	ref	ref
	Public sector	1.04	1.55***	1.16
	Self-employed	0.10***	3.57***	1.01
The number of cases (n)		1037	923	930

More importantly, we analyzed the outcome lifetime with a finite mixture technique and discovered that it consists of four distinct sub-populations with different level of mortalities. In the application considered here, the frailty model with individual random effects yields estimates approximately the same as in the basic Cox PH model. Clearly Cox model estimates change when adding mixture component as a factor. Thus, the results show some discrepancy in parameter estimates, or bias, in basic Cox PH model estimates. Both goodness-of-the fit statistics and likelihood-ratio tests improve using the extensions provided here. Nonetheless, the frailty model yields estimates that are close to the basic Cox model. The results for Cox-mixture analysis show differences between groups regarding gender and level of education. Results on disability pension are similar compared to other models. It is likely that the group-composition, which is different compared to ex ante groups, affects the estimates.

These analyzes indicate substantially that, at this stage of the life course, there are significant differences in mortality between men and women, and indeed the expected lifetime for women of the studied cohort is known to be about two years longer than that for men. The eligibility rules of disability pension are strict because there must be a severe and long-term illnesses in order to get a pension. Therefore, the increased likelihood of premature death indicated by the models is no surprise.

Our ongoing work includes analytically merging the Cox PH model with mixture modelling techniques, with an eye to developing a much-needed open-source (*e.g.*, R) package to facilitate use by practitioners and statistical modellers. As noted previously, the cac function is designed to this kind of analysis and it can be useful in analyzing Cox mixture models.

Acknowledgements

The authors wish to thank the anonymous referee for many thoughtful comments that greatly improved this paper.

References

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V. and Walther, G. (2007). Advances in mixture models. *Computational Statistics and Data Analysis*, **51**, 5205–5210.

Collett, D. (2015). *Modelling Survival Data in Medical Research.* Third Edition. CRC Press, Boca Raton, FL.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society,* **B34**, 187–220.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, **B39**, 1–38.

Eng, K. H. and Hanlon, B. M. (2014). Discrete mixture modeling to address genetic heterogeneity in time-to-event regression. *Bioinformatics*, **30**, 1690–1697.

Ha, I. D., Jeong, J.-H. and Lee, Y. (2017). *Statistical Modelling of Survival Data with Random Effects. H-Likelihood Approach.* Springer Nature, Singapore.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Heckman, J. J. and Singer, B. S. (Eds.) (1985). *Longitudinal Analysis of Labor Market Data.* Cambridge UK Econometric Society Monographs 10. Cambridge University Press, Cambridge, UK.

Machin, D., Cheung, Y. B. and Parmar, M. (2006). *Survival Analysis: A Practical Approach.* Second Edition. Wiley, Chichester.

McLachlan, G. J. and McGiffin, D. C. (1994). On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research*, **3**, 211–226.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley, Hoboken, NJ.

Nagin, D. S. (2005). *Group-Based Modeling of Development.* Harvard University Press, Cambridge, MA.

Nagin, D. S. and Odgers, C. L. (2010a). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology*, **6**, 109–138.

Nagin, D. S. and Odgers, C. L. (2010b). Group-based trajectory modeling (nearly) two decades later. *Journal of Quantitative Criminology*, **26**, 445–453.

Ng, S. K., Xiang, L. and Yau, K. K. W. (2019). *Mixture Modelling for Medical and Health Sciences.* Chapman & Hall/CRC, New York.

Nummi, T., Salonen, J. and O'Brien, T. (2017). Statistical analysis of labor market integration: A mixture regression approach. In D.-G. Chen, Z. Jin, G. Li, Y. Li, A. Liu and Y. Zhao, Editors, *New Advances in Statistics and Data Science*, ICSA Book Series in Statistics, 313–321. Springer, Cham.

Official Statistics of Finland (2020). Population projection (e-publication). Helsinki: Statistics Finland [Referred: 28.2.2020]. http://www.stat.fi/til/vaenn/tau_en.html.

Polvinen, A., Laaksonen, M., Gould, R., Lahelma, E., Leinonen, T. and Martikainen, P. (2015). Socioeconomic inequalities in cause-specific mortality after disability retirement due to different diseases. *Scandinavian Journal of Public Health*, **43**, 159–168.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sanchez, L. B., Lachos, V. H. and Moreno, E. J. L. (2018). *CensMixReg: Censored Linear Mixture Regression Models.* Version 3.1. https://cran.r-project.org/src/contrib/Archive/CensMixReg/.

Sewdas, R., de Wind, A., Stenholm, S., Coenen, P., Louwerse, I., Boot, C. and van der Beek, A. (2020). Association between retirement and mortality: Working longer, living longer? a systematic review and meta- analysis. *Journal of Epidemiology and Community Health*, **0**, 1–8.

Therneau, T. M. (2020). *coxme: Mixed Effects Cox Models.* Version 2.2-16. https://CRAN.R-project.org/package=coxme.

Therneau, T. M. and Lumley, T. (2019). *Survival: A Package for Survival Analysis in S.* Version 3.1-8. https://CRAN.R-project.org/package=survival.

van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations. In J.J. Heckman and E.E. Leamer, (Eds.), *Handbook of Econometrics*, First Edition, Volume **5**, Chapter **55**, 3381–3460. Elsevier.

Wienke, A. (2010). *Frailty Models in Survival Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

Zhang, Y. (2008). *Parametric Mixture Models in Survival Analysis with Applications.* Ph.D. thesis. Temple University.

APPENDIX

Figures

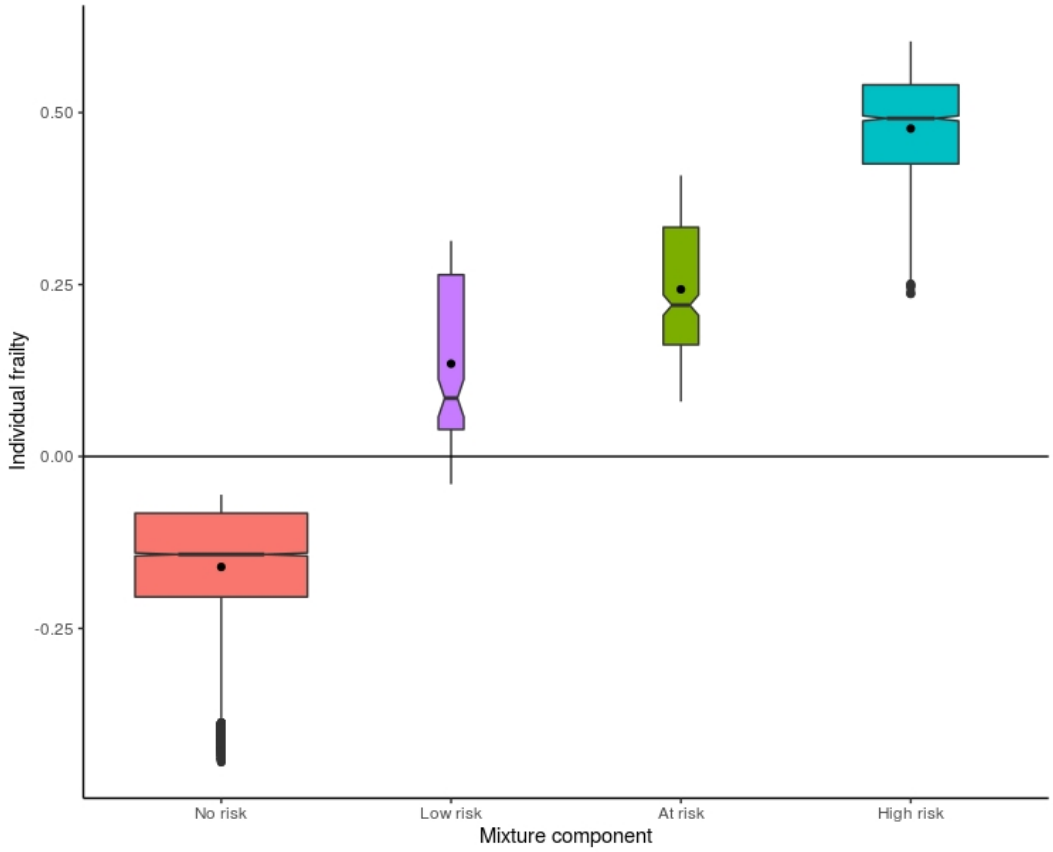

Figure 3: Box plot of individual frailty coefficients by mixture component. Box widths are proportional to sample sizes. The notches indicate the sample medians and the dots indicate the sample means.

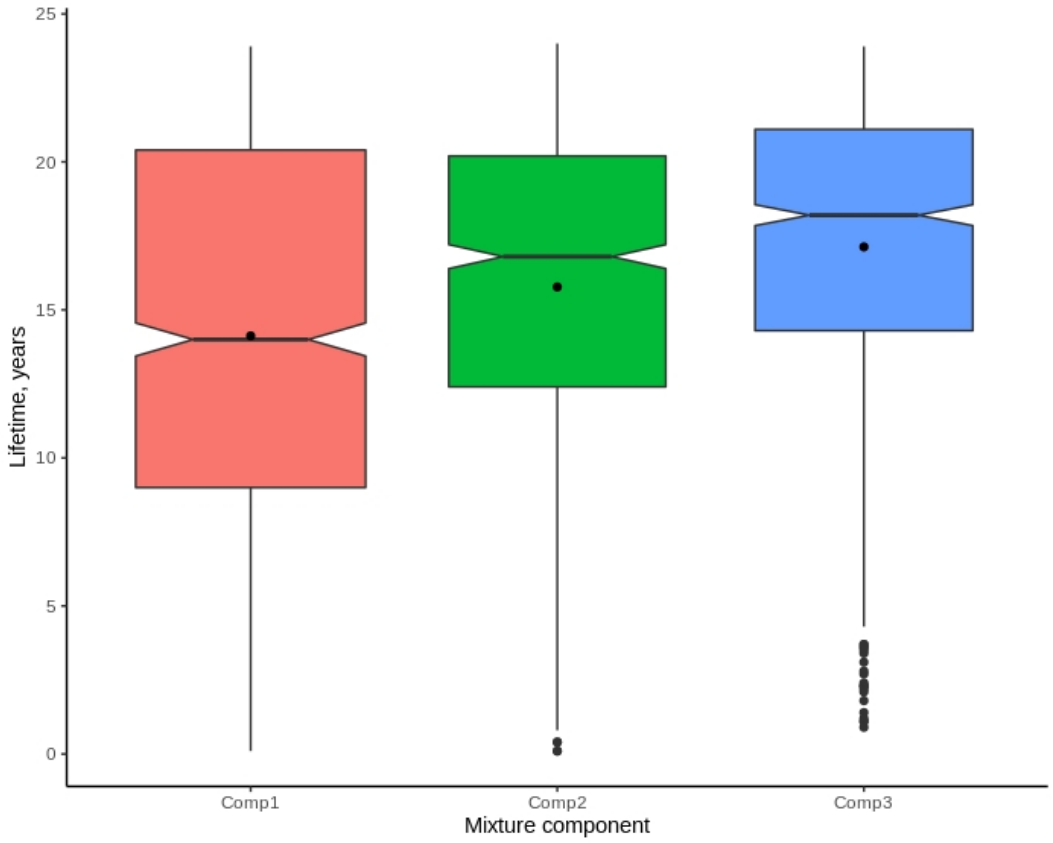

Figure 4: Box plot of lifetimes in mixture Cox model by mixture component. Box widths are proportional to sample sizes. The notches indicate the sample medians and the dots indicate the sample means.