Mixture of Cox regression models with L 1 -penalization for modeling patients survival time after liver transplantation

Eliz Peyraud, Julien Jacques, Guillaume Metzler, Ines Faivre, Mathis Dousse

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Mixture of Cox regression models with L_1-penalization for modeling patients survival time after liver transplantation

Eliz Peyraud[1,2] | **Julien Jacques**[1] | **Guillaume Metzler**[1] | **Ines Faivre**[1] | **Mathis Dousse**[1]

[1]Université Lumière Lyon 2, Laboratoire ERIC UR 3083, Lyon, France

[2]I-cair, Institute George Lopez, Lyon, France

Correspondence
Corresponding author Eliz Peyraud.
Email: epeyraud@igl-transplantation.com

Abstract

The study of time-to-event data is essential in fields such as biology and medicine, particularly for improving patient outcomes. For example it enables the evaluation of survival times in cancer patients or the duration until a relapse occurs. In liver transplantation, which is crucial for treating end-stage liver diseases, the increasing demand for grafts and the limited availability of donors have led to the use of extended criteria donors. This approach, while addressing the graft shortage, also increases the risk of graft failure. Traditional survival models are often insufficient for these high-risk scenarios and as a result, there is a critical need for highly advanced models that can reliably predict the success or failure of transplantations under these increasingly complex conditions. In parallel, a significant amount of data collected over recent years offers valuable insights into graft failures and patient survival. This wealth of data has generated numerous indicators, yet only a limited number are utilized in practice. Moreover, the diversity within the transplant patient population has introduced a significant heterogeneity in the data, which must be carefully managed to ensure accurate analysis and application. To effectively utilize this data, advanced statistical tools are needed. The Cox proportional hazards model, a well-established method in survival analysis, can be instrumental in this regard. This model and its extensions, such as mixture models or penalization techniques, provide robust frameworks for predicting patient outcomes and understanding the factors influencing survival.

This article proposes a Deep Penalized Cox Mixture model (DPCM). The application of such a mixture of L_1-penalized Cox model to liver transplant data will enable us to simultaneously establish patient subgroups with similar behavior and select the most relevant variables for determining survival time from the extensive data available today. This approach not only enhances our understanding of the factors contributing to graft failure but also improves the prediction of survival time for each patient within the entire population, ultimately leading to better outcomes in liver transplantation. Moreover, our model delivers better results compared to existing approaches that utilize only the mixture part or the penalty part, offering a more effective solution. This approach aims to enhance decision-making and improve patient care in the context of liver transplantation.

KEYWORDS
Proportional hazard model, mixture modelling, penalization, liver data

1 | INTRODUCTION AND MOTIVATION

The study of time-to-event data is nowadays essential to the development of scientific disciplines such as medicine. In this field, where predicting outcomes over time is vital, such analyses are crucial for improving patient care and treatment strategies. Time-to-event data is indispensable for evaluating patient survival rates, or treatment effectiveness, ultimately guiding clinical decision-making and improving outcomes. In this context, our focus will be on liver transplantation, which is currently the only treatment to cure end-stage liver diseases.

Some prognosis scores were developed to estimate patient survival and to sort and prioritize patients on the waiting list such as the MELD/PELD score (Model for End-Stage Liver Diseases/ Pediatric End-Stage Liver Diseases)[1], the D-MELD score (Donor Age - Model for End-Stage Liver Diseases)[2] and the DRI score (Donor risk index)[3] to name but three.

For example, the MELD score[1] is calculated using the values of serum bilirubin, serum creatinine, and INR (a measure of blood coagulation) and is given by the formula: $R = 9.57 \times \log(\text{creatinine mg/dL}) + 3.78 \times \log(\text{total bilirubin mg/dL}) + 11.2 \times \log(\text{INR}) + 6.43$. These values —serum bilirubin, serum creatinine, and INR— are recognized as crucial indicators of liver function. Elevated bilirubin reflects impaired liver processing, while an increased INR indicates reduced production of clotting factors by the liver. Although creatinine primarily measures kidney function, its inclusion is relevant due to the impact of liver disease on the kidneys. While these three indicators are essential for the MELD score, advances in medical diagnostics now allow for the evaluation of additional factors that could provide a more complete picture of liver disease. Liver allocation is currently based on the MELD score, since the United Network for Organ Sharing (UNOS) ‡ adopted and approved the MELD score to allocate organs for patients awaiting liver transplantation in the United States in 2002.

However, for several years, the demand for transplants has been steadily increasing, while the number of available donors has not kept pace proportionally, creating a shortage of grafts. This recent shortage of transplants has led to the use of expanded-criteria donors[4]. To increase the number of donors, surgeons are resorting to the use of grafts from expanded criteria donors, however this approach carries a higher risk of graft failure, and the models currently used in practice to evaluate patient survival are not well-suited for these cases. This underscores the need for the creation of more performant models capable of better predicting outcomes in such high-risk scenarios.

A notable advancement is the extensive accumulation of data over recent years, which is now available to enhance our understanding of graft failures and to develop predictive models for patient survival using advanced statistical tools. Despite this wealth of information, physicians and surgeons cannot utilize all the data real-time during transplantation procedures. Therefore, a critical selection process is required to determine which data are most relevant for clinical decision-making. This selection process can be significantly improved through the development of new survival models, which will aid physicians in comprehending the patient populations they treat and identifying key factors that influence survival outcomes.

Indeed, the main challenge of survival analysis is to estimate and understand the time until the occurrence of an event of interest, such as the survival time of a patient following surgery. There are numerous survival analysis methods available today such as the Kaplan-Meier estimator[5], random survival forests[6], and parametric survival models[7]. Among all of them, Cox proportional hazards model[8] is the most frequently used survival estimation method for problems involving the presence of covariates. This is explained by the robustness, flexibility and the explanability aspect of the model where the impact of the covariates on the survival probability can be measured. Introduced by Cox D.R. in 1972[8], this semi-parametric model, estimated by partial likelihood maximization[9], allows to estimate the hazard risk of patient death. This model is particularly valued for its ability to handle censored data without assuming a specific baseline hazard function, making it highly versatile in various applications.

Since then, this model has been extensively studied, and numerous extensions have been made to improve its performance.[10].

In our research, we are particularly interested in the modeling of groups of patients that are inherently heterogeneous—due to variations in factors such as age, gender, ethnicity, physical condition, and medical history. This variability needs to be carefully considered in the modeling process. One simple method method would be to manually create subgroups of patients and apply separate models to each. However, this approach is not optimal, as it is difficult to objectively determine which criteria —such as age, medical history, or donor type— should define the subgroups. A more robust solution is to allow the model to automatically detect these subgroups, eliminating the bias of manual selection. Mixture models[11] offer a particularly powerful framework for this, as they enable the model to identify distinct subgroups and capture the complex relationships between patient characteristics and survival outcomes.

The mixture of Cox models was introduced by Rosen et al.[12]. The approach has been further developed and refined, offering a robust framework for handling heterogeneous data in survival analysis. The most recent advancements include the work by Nagpal, et al.[13], which leverages deep learning techniques to enhance the flexibility and accuracy of these models. Additionally, Ng et al.[14] provide comprehensive insights and methodologies related to mixture modeling in the medical field, offering a thorough foundation for applying these models in practice.

The second challenge identified in our study on liver transplantation is managing the extensive amount of patient data available. In clinical practice, it is neither feasible nor realistic for physicians to consider all variables in real-time decision-making, and

‡ https://unos.org/

including too many variables in a model can lead to overfitting, which compromises the accuracy of the results[15]. To address this issue, we focus on reducing the dimensionality of the data while ensuring that the model remains interpretable for clinicians. This is achieved through the application of an L_1 penalty, which forces some model parameters to zero when they are not significant, allowing us to reduce the dimensionality without sacrificing the model interpretability.

The penalized Cox model, incorporating the L_1 penalty, represents therefore a key extension of the traditional Cox model, particularly following the development of LASSO regression. Tibshirani's seminal paper[16] introduced the application of the L_1 penalty to the likelihood function of the Cox model. Since then, new methods for estimating the penalized Cox model have emerged, incorporating more advanced optimization algorithms. Notable contributions include the optimization improvements by Goeman[17] and Simon et al.[18]. These developments significantly enhance the practicality and effectiveness of penalized Cox models in handling high-dimensional data, making them particularly valuable for our application.

To further improve modeling, there is a need to simultaneously identify patient subgroups and determine the most relevant features for these groups. Integrating the mixture of Cox models with an L_1-penalty addresses this need by uncover distinct patient groups and select important features. Such a model holds great potential for real-world applications by addressing multiple challenges simultaneously and represents a significant advancement, as demonstrated by Webb et al.[19], who explored the benefits of combining these methodologies. However, their work was limited to only two subgroups and fixed the number of components, which highlights the potential for further exploration and flexibility in applying these methods. This limitation presents an opportunity for improvement, for which the model proposed in this article serves as a solution.

In liver transplantation, adopting such a model could have substantial clinical benefits. It could reveal previously unknown subgroups of patients who exhibit similar behaviors but are not yet identified by current medical practices. Additionally, the model could uncover new factors that influence patient survival post-transplant, providing insights beyond the existing organ allocation criteria. By better understanding these new subgroups and their characteristics, the model could enhance donor-recipient matching and improve care at transplantation centers, ultimately leading to better-targeted interventions and improved transplant outcomes.

The remainder of this paper is structured as follows: After an overview of the background on the Cox proportional hazards model and its extension to the mixture case in Section 2, our proposed inference process, named Deep Penalized Cox Mixture (DPCM) is presented in Section 3. Then, in Section 4, DPCM is compared to competitors on a set of benchmarks, and finally in Section 5 on the Scientific Registry of Transplant Recipients (SRTR) [§] database to interpret the results on liver transplantation. The paper concludes with a discussion of the findings and explores their implications as well as potential future research directions based on the results.

2 | BACKGROUND

In survival analysis, the goal is to predict the time until an event occurs, such as the survival time of patients. When we have additional information or covariates—such as age, treatment type, or biological factors—it is crucial to incorporate them into the model. To achieve this, we turn to the Cox proportional hazards model, a powerful and widely recognized tool specifically designed to assess the impact of covariates on survival time. In this section, we detail the mathematical tools required to build the Cox model (Section 2.1), followed by its extension to Cox mixture models (Section 2.2) and a non-linear variant (Section 2.3). These concepts are essential for understanding the inference of the Deep Cox Mixture model using the DPCM method presented in Section 3.

2.1 | The Cox proportional hazard regression

The Cox regression[8] is a semi-parametric model used to design the occurrence of events (such as the death of a patient) over time, as a function of covariates. The survival time (the duration from a starting time t_0 until the event of interest occurs) is modeled by the hazard function λ as:

$$\lambda(t, X) = \lambda_0(t) \exp(\boldsymbol{\beta}^T X),$$

[§] https://www.srtr.org/

where $t \in \mathbb{R}^+$ is the time, $X \in \mathbb{R}^p$ the vector of covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of regression coefficients and $\lambda_0(t)$ is the so-called baseline hazard function (i.e. the hazard function of event when all covariates are zero). In the semi-parametric setting, no specific form is assumed on $\lambda_0(t)$.

The hazard function λ is directly linked to the survival function S, which represents the probability that an individual survives beyond time t, with the following relation:

$$f(t|X) = \lambda(t, X)S(t|X),$$

where $f(t|X)$ is the probability density function of the random variable T, which represents the time until the event occurs for an individual with covariates X.

Moreover, survival data can be subject to right-censoring. Censoring occurs when the exact date of death is unknown. In such cases, the date of the last follow-up is used to calculate the survival time instead of the date of death, indicating that the patient was known to be alive up until that date, after which information is no longer available. We denote by δ_i the censoring indicator associated with each individual's survival time ($\delta_i = 0$ if the individual is censored, 1 otherwise).

Based on the observation of a sample $(X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)$, the estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the log-likelihood ℓ:

$$\ell(\boldsymbol{\beta}, X_1, \ldots, X_n, \delta_1, \ldots, \delta_n) = \sum_{i=1}^{n} f(t_i|X)^{\delta_i} S(t_i|X_i)^{1-\delta_i},$$

corresponding respectively to the the survival function for censored individuals ($\delta_i = 0$) and the density function for uncensored individuals ($\delta_i = 1$).

Once the parameter of the hazard risk function have been estimated, the probability of survival for the individuals can easily be calculated by the relationship[20]:

$$S(t|X) = \exp\left(-\exp(\boldsymbol{\beta}^T X) \int_0^t \lambda_0(u|X)\, du\right).$$

Employing a single model for a population assumes homogeneity across individuals. Due to this assumption, the practice is typically to pre-select homogeneous patient groups (based on criteria such as age or medical history). One way of overcoming these difficulties is to use a mixture of Cox models[14]. This method accounts for patient heterogeneity by creating population subgroups, and then the components of the mixture can be analyzed a posteriori to identify new criteria that define these homogeneous groups.

2.2 | Mixture of parametric Cox models

To study heterogeneity within individuals and identify population subgroups, a commonly used method is mixture modeling[14]. We focus on a mixture with $g \in \mathbb{N}$ components. The survival function is expressed as the sum of the survival function S_h of each component mixture multiplied by their respective proportions π_h ($h = 1, \ldots, g$):

$$S(t|X) = \sum_{h=1}^{g} \pi_h S_h(t|X).$$

The same applies to the density function f of the random variable T, expressed as the sum of the probability density function f_h of each component mixture multiplied by their respective proportions π_h ($h = 1, \ldots, g$) :

$$f(t|X) = \sum_{h=1}^{g} \pi_h f_h(t|X).$$

Since the hazard function is defined as the ratio between the density function f and the survival function S, the shapes of the density functions of the mixture are easily found by the relation $f_h(t|Xt) = \lambda_h(t, X)S_h(t|X)$.

As mentioned in Section 2.1, the baseline hazard function is generally unknown in the semi-parametric Cox model. However, in certain cases, we can assume a specific form for λ_0, based on a particular distribution or functional form. In this study, focusing on patient lifetime data, we opt for a parametric approach by specifying the form of λ_0. This choice simplifies the model inference process, especially when dealing with mixture models, and allows for more straightforward estimation.

Assuming that the baseline hazard follows a Weibull distribution, the function can then be written as:

$$\lambda_0(t) = l\alpha t^{\alpha-1},$$

with $l, \alpha > 0$.

Let Θ be the whole set of parameters to be estimated: $\Theta = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_g^T, \alpha_1^T, \ldots, \alpha_g^T, l_1^T, \ldots, l_g^T,)$. The choice of the form of λ_0 is not restrictive since it allows us, with a Weibull distribution, to be for the baseline function either increasing (when $\alpha > 1$) or monotomous (when $\alpha = 1$) or decreasing (when $\alpha < 1$) according to the value of α. In this case, it allows the real risk of patients after surgery to be correctly represented: Indeed the hazard function can be segmented into three overlapping temporal phases. Initially, there is an early phase immediately following treatment where the risk is relatively high but gradually declines. This is followed by a middle phase where the risk remains constant over time. Finally, a late phase occurs, during which the risk of failure begins to rise slowly as the patient ages[14].

2.3 | Non-linear case

In the Cox model, the logarithm of the hazard function is explained as a linear function of the covariates. However, a non-linear variant[21] offers the potential to capture more complex relationships between predictors and survival outcomes. This non-linear version enhances the model's flexibility by introducing non-linear effects, allowing for a richer representation of the data that may uncover patterns overlooked by linear models. The goal is to provide a deeper understanding of survival dynamics by extending beyond the constraints of linearity, improving model performance and flexibility.

To capture non-linear relationships between the covariates and the survival outcome, the model[21] uses transformations of the covariates. For example, terms such as polynomial functions may be included. Let $\phi(\boldsymbol{\beta}^T X_i)$ denote a non-linear transformation of $\boldsymbol{\beta}^T X_i$. The hazard function in this case can be expressed as:

$$\lambda(t, X_i) = \lambda_0(t) \exp\left(\phi(\boldsymbol{\beta}^T X_i)\right).$$

The non-linear model used is the Deep Cox Mixture (DCM), which assumes a mixture of non-linear Cox models, integrates a neural network to capture complex relationships in survival data. The introduction of non-linear effects leads to a loss of interpretability, which may pose challenges when applying the model to real-world data. Understanding and interpreting the results of non-linear models can be more delicate, requiring careful consideration of how these advanced techniques influence the insights derived from the data.

3 | INFERENCE OF THE DEEP COX MIXTURE MODEL

This section focuses on the inference of the linear model, drawing on methodologies utilized for the inference of the mixture model[14] and the penalized model[17]. By integrating concepts from both the mixture and penalized models, the DPCM inference of the Deep Cox Mixture model will be presented. This approach represents a significant contribution to the field, as it simultaneously addresses the challenges of clustering and variable selection within survival analysis. Given the clinical imperative to tailor models for specific patient subgroups predictions while maintaining interpretability, thereby meeting a critical need in medical research and practice (for example it is recognized that certain patients need customized models[22] based on factors such as age and donor type; however, the specific variables required for accurately predicting survival in these patients are still unclear).

The likelihood L of the mixture model (in the parametric case) can be rewritten as follows:

$$L(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, t_i, \boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha})^{\delta_j} S(\mathbf{x}_i, t_i, \boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha})^{1-\delta_j}, \tag{1}$$

$$= \prod_{i=1}^{n} \left(\sum_{h=1}^{g} \pi_h f_h(\mathbf{x}_i, t_i, \beta_h, \mathbf{l}_h, \boldsymbol{\alpha}_h) \right)^{\delta_j} \left(\sum_{h=1}^{g} \pi_h S_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, \mathbf{l}_h, \boldsymbol{\alpha}_h) \right)^{1-\delta_j}, \tag{2}$$

where $(\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = (\mathbf{x}_1, \ldots, \mathbf{x}_n, t_1, \ldots, t_n, \delta_1, \ldots, \delta_n)$ represents the set of individuals with their respective lifetimes and censoring indicators and $(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}) = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g, l_1, \ldots, l_g, \alpha_1, \ldots, \alpha_g)$ the set of parameters according to each component of the mixture model.

In high dimensional survival analysis, the standard Cox proportional hazards model can encounter challenges when the number of covariates is large, even if it does not exceed the number of individuals. To enhance model performance and facilitate variable

selection, penalization techniques, such as the L_1-penalty [16], are employed. The penalized Cox model [16] improves interpretability and reduces the risk of overfitting by focusing on a manageable number of covariates.

Dimension reduction is also crucial for this study, as it addresses the abundance of variables and enhances model interpretability. Given that this work is aimed at a clinical and medical audience, maintaining the interpretability of the results is essential. Thus, dimension reduction methods that alter the nature of the variables, such as PCA, are not appropriate. Instead, the L_1-penalty is preferable to select the most relevant variables for survival estimation.

When the number of covariates is large, we aim to reduce their number while maintaining the interpretability of the results.To do this, following the approach of Goeman [17], a L_1-penalty is applied to the log-likelihood function of the Cox mixture model. The penalized log-likelihood ℓ_{pen} can then be written as follows:

$$\ell_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \sum_{i=1}^{n}[\delta_i \log(\sum_{h=1}^{g} \pi_h f_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h)) + (1-\delta_i)\log(\sum_{h=1}^{g} \pi_h S_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h))] - \eta \sum_{h=1}^{g}\sum_{j=1}^{p} |\boldsymbol{\beta}_h^j|. \quad (3)$$

where the strength of the penalty is determined by the $\eta \geq 0$ parameter.

Note that a decision was made to apply a single penalty term across all components of the mixture model. This means that the penalty term will have the same value for each component, regardless of differences between them. This choice was made to reduce model complexity and will be discussed in Section 6.

The model parameters $\Theta = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_g^T, \alpha_1^T, \ldots, \alpha_g^T, l_1^T, \ldots, l_g^T,)$ are estimated by maximizing the penalized log-likelihood (3).

Given our focus on the parametric case, all the model parameters are estimated by maximum likelihood using an Expectation-Maximization (EM) algorithm, as the presence of a mixture makes direct analytical maximization impossible. Therefore the EM numerical algorithm is particularly well-suited for this context. The EM algorithm iteratively alternates the two following steps: the E-step, which involves calculating the conditional expectation of the complete likelihood $\ell_{c,\text{pen}}$ given the observed data, and the M-step, which maximizes this conditional expectation with respect to the model parameters. This process is repeated until the likelihood converges.

A latent binary variable Z_{ih} is introduced, where $Z_{ih} = 1$ if individual i belongs to class h, and 0 otherwise. The penalized complete log-likelihood of the Deep Cox Mixture model can then be written as follows:

$$\ell_{c,\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}, Z) = \sum_{i=1}^{n}\sum_{h=1}^{g} Z_{ih}[\delta_i \log(\pi_h f_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h)) + (1-\delta_i)\log(\pi_h S_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h))] - \eta \sum_{h=1}^{g}\sum_{j=1}^{p} |\boldsymbol{\beta}_h^j|. \quad (4)$$

(**E-step**) Let $\mathbb{E}(Z_{ih}|\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \tau_{i_h}$, representing the conditional expectation of Z_{ih}. The E-step consists of calculating the probability for an individual i to belong to the h component of the mixture model, as part of the conditional expectation of the penalized complete likelihood $\ell_{c,\text{pen}}$. The formula for τ_{i_h} can be expressed as follows:

$$\tau_{i_h} = \frac{(\mathbb{1}_{\{\delta=1\}} f_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h) + \mathbb{1}_{\{\delta=0\}} S_h(\mathbf{x}_i, t_i, \boldsymbol{\beta}_h, l_h, \alpha_h)) \times \pi_h}{\sum_{l=1}^{g}(\mathbb{1}_{\{\delta=1\}} f_l(\mathbf{x}_i, t_i, \boldsymbol{\beta}_l, l_l, \alpha_l) + \mathbb{1}_{\{\delta=0\}} S_l(\mathbf{x}_i, t_i, \boldsymbol{\beta}_l, l_l, \alpha_l)) \times \pi_l}. \quad (5)$$

(**M-step**) Updating the parameters $(\pi_1, \ldots, \pi_{g-1})$ at iteration k is classically done by:

$$\pi_h^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_{ih}^{(k)}}{n}, \qquad \forall h \in \{1, \ldots, g\}, \quad (6)$$

Due to the presence of the absolute value of β_h^j in Formula (3) the likelihood presented here is not differentiable at all points. In contrast, the other parameters, α and l, are not affected by this issue. Consequently, it becomes necessary to define a directional derivative [17] specifically for β_h^j. The directional derivative is written as follow:

$$\ell'_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}; \mathbf{v}) = \lim_{t \to 0} \frac{1}{t}((\ell_{\text{pen}}(\boldsymbol{\beta} + t\mathbf{v}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) - (\ell_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta})), \qquad \forall \mathbf{v} \in \mathbb{R}^p.$$

Let us note $\mathbf{v}_{\text{opt}}$ the optimal direction that maximizes $\ell'_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}; \mathbf{v})$, then the gradient $\mathbf{g}$ can be define as:

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{cases} \ell'_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}; \mathbf{v}_{\text{opt}})\mathbf{v}_{\text{opt}} \text{ if } \ell'_{\text{pen}}(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}; \mathbf{v}_{\text{opt}}) \geq 0, \\ 0 \text{ otherwise.} \end{cases}$$

Finally, the gradient $\mathbf{g}(\boldsymbol{\beta})$ can be define for each component $h = 1, \ldots, g$ of the mixture from the unpenalized log-likelihood gradient $\mathbf{h}(\boldsymbol{\beta})$ such as:

$$\mathbf{g}^j(\boldsymbol{\beta}_h) = \begin{cases} \mathbf{h}^j(\boldsymbol{\beta}_h) - \eta \text{sign}(\boldsymbol{\beta}_h^j) \text{ if } \boldsymbol{\beta}_h^j \neq 0, \\ \mathbf{h}^j(\boldsymbol{\beta}_h) - \eta \text{sign}(h^j(\boldsymbol{\beta}_h)) \text{ if } \boldsymbol{\beta}_h^j = 0 \text{ and } |h^j(\boldsymbol{\beta}_h)| > \eta, \\ 0 \text{ otherwise.} \end{cases},$$

where $\mathbf{h}^j(\boldsymbol{\beta}_h) = \dfrac{\partial \ell(\boldsymbol{\beta}_h^j, \mathbf{l}_h, \boldsymbol{\alpha}_h, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}_h^j}$.

The update is so performed using the gradient ascent algorithm on each component of the mixture during the M-step of the EM algorithm. The update for $\boldsymbol{\beta}_h$ at step $k+1$ is given by:

$$\boldsymbol{\beta}_h^{(k+1)} = \boldsymbol{\beta}_h + \rho \, \mathbf{g}(\boldsymbol{\beta}_h^{(k)})$$

where $\rho \in \mathbb{R}$ is the learning rate of the gradient ascent (this parameter controls the step size during updates and can significantly affect convergence speed and stability. It is typically chosen through tuning methods, making it a hyperparameter of the method).

When it comes to the parameters $(\alpha_h^{(k+1)}, l_h^{(k+1)})$, the update is done by maximizing the log-likelihood function, i.e. by cancelling its gradient. After calculation and in the case where the baseline function follows a Weibull distribution, the new values of $(\alpha_h^{(k+1)}, l_h^{(k+1)})$ are obtained by solving the following non-linear system:

$$\begin{cases} \sum_{i=1}^n \tau_{ih}^{(k)} \left(\dfrac{\delta_i}{l_h^{(k)}} - \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_h^{(k+1)}\right) l_h^{(k)} t_i^{\alpha_h^{(k)}} \right) = 0, \\ \sum_{i=1}^n \tau_{ih}^{(k)} \left(\dfrac{\delta_i}{\alpha_h^{(k)}} + \delta_i \log(t_i) - \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}_h^{(k+1)}\right) l_h^{(k)} t_i^{\alpha_h^{(k)}} \log(t_i) \right) = 0. \end{cases}$$

This non-linear system does not have an analytic solution, so in practice, numerical optimization algorithms are employed to find an approximate solution.

The algorithm stopping criterion is defined as the relative change in the penalized log-likelihood being less than a predefined threshold, typically set at 0.001. Initialization is performed using random values for each parameter of the model. Convergence properties explain the necessity of multiple initializations, as the EM algorithm can converge to different local optima based on initial values. To mitigate this, the algorithm is run multiple times with different initializations values (ypically run 10 to 20 times). The number of groups is chosen based on the optimization of the selected evaluation metric (specific details on the evaluation metric provided in Section 4).

The same DPCM inference technique with the EM algorithm can be applied to non-linear Cox regression, as it follows a similar approach to the linear case, with the primary difference being the last system of equations that needs to be solved. In the non-linear case, the system is adapted to accommodate the additional complexity introduced by non-linear transformations, and we employ a neural network to model these non-linear effects. Despite these changes, the same type of numerical algorithm is used to optimize the parameters, ensuring consistency in the inference process.

4 | APPLICATIONS

The Deep Penalized Cox Mixture method presented in the previous section enables innovative and promising studies to be carried out in the medical field. In this section, the first part will be devoted to the evaluation of the model on a benchmark of known, open-access health datasets, while the second part will focus on the analysis of the model results on data provided by the SRTR and targeting patients who have undergone liver transplantation in the past few years. The implementation of the inference algorithm was done by extending the code developed by Nagpal et al. for the Deep Cox model [13]. It is important to note that the available code for the Deep Cox model [13] provides an implementation where both linear effects (as in the classic Cox model) and non-linear effects (corresponding to their DeepCox model) can be considered. Starting from their code, we modified it in order to incorporate an L_1 penalization into the inference.

Protocol

To evaluate the performance of the DPCM method across various datasets, we performed 10 estimations of the model parameters for each configuration. The dataset was split into 60% for training, 20% for validation, and 20% for testing. Four models are evaluated:

1. **CPH** : Standard Cox proportional hazards model (i.e., with a single mixing component and no penalty)[8].
2. **Mixture CPH** : Mixture of Cox models[14] (without penalization), with k the number of components $h \in \{2, 3\}$ (will be extended to 5 for application to liver transplantation).
3. **Penalized CPH** : Penalized Cox model[17] (without mixture) with the penalty hyperparameter η acrossing the following values: $\eta = \{0, 0.001, 0.1, 0.5, 1, 10, 100\}$.
4. **DPCM** : Our proposed Deep Penalized Cox Mixtures inference technique, with variations in both the number of components k and the penalty hyperparameter α in the same range as above.

For all models, the learning rate was tuned between the two values: $\{1e-3, 1e-4\}$.

Additionally, we extended the evaluation of the four previous model cited above to a non-linear framework as mentionned in Section 2.3 to assess its robustness and effectiveness. For this purpose, we employed the same methodology as described above, with the inclusion of a neural network layer. The neural network configurations included two variations of the number of layers: (i) with a single hidden layer containing 100 units, (ii) with two hidden layers, each containing 100 units. The learning rate for these models remained consistent with the aforementioned values.

Evaluation metric

To evaluate the model's performance on each of these datasets, the integrated Brier score[23] was used as our metric. Indeed, the Brier score allows to evaluate the accuracy of a predicted survival function at a given time t; it represents the average squared distances between the observed survival status and the predicted survival probability. It is an extension of the mean squared error to censored data, whose formula is shown below:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{1}(t_i > t) - S(t, \mathbf{x}_i))^2$$

where $\mathbf{1}(t_i > t)$ is the indicator function taking a value of 1 if the individual i is still alive at time t (i.e. $(t_i > t)$) and 0 otherwise.

The Integrated Brier Score (IBS) is an extension of the Brier score, which is calculated by integrating the time-dependent Brier Score $BS(t)$ over the interval $[t_1; t_{max}]$ where t_1 is the starting time of interest (usually 0) and t_max s the maximum follow-up time, and using the weighting function $w(t) = \frac{t}{t_{max}}$:

$$\text{IBS} = \int_{t_1}^{t_{max}} \text{BS}(t)dw(t) = \frac{1}{t_{max}}\int_{t_1}^{t_{max}} \text{BS}(t)d(t) \tag{7}$$

The integral in Formula (7) typically lacks an analytical solution and is therefore usually computed through numerical approximation methods, such as Simpson's rule. Thus, minimizing the IBS leads to optimize the model selection. Additionally, it provides the selection rates of variables for penalized methods and the number of components for mixture models.

4.1 | Benchmark study

This section introduces the benchmark datasets used to evaluate the performance of the different models cited above. These datasets provide a diverse and robust foundation for assessing the model's accuracy and reliability. By benchmarking our approach against these datasets, we aim to demonstrate its effectiveness in various scenarios and establish a standard for comparison with other models.

4.1.1 | Datasets presentation

Three benchmark data sets are considered:

- **PBC dataset**: This data set come from a Mayo Clinic trial conducted between 1974 and 1984, involving 424 patients with Primary sclerosing cholangitis (PSC) who participated in a randomized, placebo-controlled trial of the drug D-penicillamine. PSC is an autoimmune disease that progressively destroys the small bile ducts in the liver, leading to cirrhosis and liver decompensation over time. The PBC dataset consists of 1,945 individuals and 20 variables.[¶]
- **Framingham dataset**: The Framingham Study aims to investigate the incidence and prevalence of CardioVascular Disease (CVD) and its risk factors, trends over time, and familial patterns. Additionally, it seeks to estimate disease incidence rates and describe the natural history of CVD, including the sequence of clinical signs before it becomes clinically recognizable and the progression of the disease once it manifests. The data set includes information from the first 32 clinical exams, selected ancillary data, and event follow-up through 2018. The Framingham dataset consists of 11,627 individuals and 29 variables.[#]
- **SUPPORT dataset**: This dataset originates from a Vanderbilt University study aimed at estimating survival for seriously ill hospitalized adults. It is a random sample of patients drawn from Phases I & II of the SUPPORT study (Study to Understand Prognoses, Preferences, Outcomes, and Risks of Treatment). The SUPPORT dataset consists of 9,105 individuals and 24 continuons and categorical variables. After binarization of the categorical variables, the total number of variables increases to 38.[‖]

4.1.2 | Results

Table 1 presents the IBS results (average and standard deviation over the 10 training/validation/test sampling) for the four models cited in Section 4.

As we are currently working the estimation of a penalized mixture of Cox models, it involves significant computational costs. Detailed running times for the experiments conducted are therefore presented in Appendix A.

Regarding the results on the PBC dataset, the impact of penalization on the IBS score stands out as the most significant among the experiments. This can likely be explained by the small number of patients in this dataset, making dimension reduction a particularly important step. The limited sample size necessitates careful variable selection, which is effectively achieved through penalization, leading to improved model performance in this context.

Examining the results in Table 1, we observe that the DPCM method consistently achieved better IBS scores across all datasets: for example with the PBC dataset, in the standard Cox model the IBS score was 0.1605 ± 0.0297, and with penalization this improved to 0.1520 ± 0.0220. Further enhancements were observed with the addition of penalization and mixture in the linear model, resulting in scores of 0.1428 ± 0.0198. When it comes to a non-linear framework using neural network layers, the performance significantly improved, with scores values of 0.1373 ± 0.0251, 0.1335 ± 0.0254, and 0.1254 ± 0.0199, respectively. Using the standard Cox model as a baseline, the results clearly demonstrate the importance of incorporating both penalization and clustering techniques to enhance the model performance.

In contrast, the limited impact of penalization on the IBS score for the Framingham dataset can be attributed to the relatively small number of variables compared to the large number of individuals. This also explains why the variable selection percentage is nearly 100%. In this context, the addition of clustering appears to have a more significant effect. The IBS score improves from 0.1026 ± 0.0033 in the linear baseline Cox model to 0.1020 ± 0.0034 with penalized mixture CPH. Further enhancement is observed in the non-linear models, where the score improves from 0.1005 ± 0.0032 to 0.0998 ± 0.0033, demonstrating the significant impact of clustering in this context. Finally, the results on the SUPPORT dataset are particularly interesting, as they demonstrate the model's effectiveness in handling categorical data. The dataset includes six categorical variables that were one-hot encoded, and the results show that both penalization and clustering contribute to improvements in the IBS score. This indicates that the model performs well even with categorical predictors, highlighting its versatility across different types of data.

5 | APPLICATION TO LIVER TRANSPLANT DATA

Following the results of our benchmark study, which examined various models using real-world data, attention is now directed toward to a more specific application. This section examines liver transplant data, focusing on its unique features and how they affect outcome modeling.

[¶] Refer to https://paperswithcode.com/dataset/pbc for the original datasource.
[#] Refer to https://clinicaltrials.gov/study/NCT00005121 for the original datasource.
[‖] Refer to http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc. for the original datasource.

Model	PBC Dataset		FRAMINGHAM Dataset		SUPPORT Dataset	
	Linear	**Non-linear**	**Linear**	**Non-linear**	**Linear**	**Non-linear**
CPH	0.1605 ± 0.0297	0.1373 ± 0.0251	0.1026 ± 0.0033	0.1005 ± 0.0032	0.1894 ± 0.0064	0.1867 ± 0.0074
Penalized CPH	0.1520 ± 0.0220	0.1335 ± 0.02546	0.1026 ± 0.0033	0.1002 ± 0.0032	0.1892 ± 0.0066	0.1860 ± 0.0074
rate var selected:	$65.55\% \pm 31.40$	$58.8\% \pm 31.93$	$99.62\% \pm 1.11$	$78.5\% \pm 20.36$	$95.78\% \pm 4.88$	$72.8\% \pm 17.43$
Mixture CPH	0.1494 ± 0.02630	0.1326 ± 0.0235	0.1021 ± 0.0034	0.1000 ± 0.0032	0.1861 ± 0.0061	0.1842 ± 0.0069
nb comp selected:	2.3 ± 0.4582	2.6 ± 0.4898	2.1 ± 0.3	2.5 ± 0.5	2.7 ± 0.4582	2.8 ± 0.3999
DPCM	$\mathbf{0.1428 \pm 0.0198}$	$\mathbf{0.1254 \pm 0.0199}$	$\mathbf{0.1020 \pm 0.0034}$	$\mathbf{0.0998 \pm 0.0033}$	$\mathbf{0.1856 \pm 0.0061}$	$\mathbf{0.1829 \pm 0.0066}$
rate var selected:	$70.27\% \pm 39.08$	$78.41\% \pm 33.77$	$99.56\% \pm 0.87$	$91.68\% \pm 11.35$	$94.69\% \pm 11.07$	$75.65\% \pm 34.84$
nb comp selected:	2.2 ± 0.39993	2.7 ± 0.4582	2.2 ± 0.39993	2.7 ± 0.4582	2.8 ± 0.3999	2.6 ± 0.4898

TABLE 1 IBS for the PBC, FRAMINGHAM, and SUPPORT Datasets: Linear and Non-Linear Cox Models, Trained 10 Times, Showing Mean and Standard Deviation. The baseline Cox Proportional Hazards Model (CPH) is compared with the Mixture of Cox Proportional Hazards Model (Mixture CPH), Penalized Cox Proportional Hazards Model (Penalized CPH), and our proposed Deep Penalized Cox Mixtures (DPCM) inference technique.

5.1 | SRTR Database

To apply the DPCM method to liver transplantation outcomes, the Scientific Registry of Transplant Recipients (SRTR)** database was utilized. The SRTR is a comprehensive database managed in the United States, collecting and analyzing data on organ transplantation nationwide. It serves as a critical resource for monitoring transplant outcomes and improving clinical practices. This study specifically focuses on liver transplantation data extracted from the SRTR, providing a robust foundation for analyzing pre- and post-transplant variables and outcomes.

This section focuses on the analysis conducted using this data, specifically examining transplantations that occurred between 2020 and 2022. This timeframe was chosen to ensure the relevance of the data by reflecting recent advancements in medical techniques and practices. The decision to focus on this period also helps in managing the size of the dataset effectively. Despite this, the selected timeframe provides valuable insights, particularly since recent studies[24,25] highlight the importance of survival rates at one, two, and three years post-transplant. Thus, even though our follow-up period extends to only four years, the data remains highly relevant to current clinical interests. The resulting dataset comprises a total of 17,064 individuals, providing a substantial sample for analysis.

In terms of variable selection, the analysis concentrated exclusively on pre-transplantation variables pertaining to the candidates, donors, and the preservation of the organs. Variables that directly provided the Model for End-Stage Liver Disease (MELD) score were excluded from the analysis to avoid redundancy and potential biases. However, the biomarkers that contribute to the calculation of the MELD score were retained to ensure that essential clinical information was preserved. This approach yielded a set of 116 continuons and categorical variables prior to encoding, which expanded to 881 variables after the encoding process was completed. Missing values in the dataset were addressed using single imputation: numerical variables were filled with the mean of the corresponding covariates, while categorical variables were replaced with the most frequent modality.

This careful selection and preparation of the dataset were undertaken to ensure the robustness and relevance of the subsequent analyses, providing a strong foundation for the exploration of the proposed models within a controlled and well-defined context.

	Linear
CPH	0.0376 ± 0.0032
Penalized CPH	0.0374 ± 0.0031 *rate var selected:* $78.58\% \pm 17.32$
Mixture CPH	0.0363 ± 0.0030 *nb comp selected:* 3 ± 0.6324
DPCM	**0.0361 ± 0.0030** *rate var selected:* $24.68\% \pm 35.25$ *nb comp selected:* 2.8 ± 1.0

TABLE 2 IBS for the SRTR Database: Linear Cox Models, Trained 10 Times, Showing Mean and Standard Deviation. The baseline Cox Proportional Hazards Model (CPH) is compared with the Mixture of Cox Proportional Hazards Model (Mixture CPH), Penalized Cox Proportional Hazards Model (Penalized CPH), and our proposed Deep Penalized Cox Mixtures inference technique (DPCM) .

5.2 | Protocol and results

The experimental protocol for this analysis follows the previously established approach, with the exception of excluding the non linear models to enhance explicability and interpretability within the clinical and medical context. Moreover, the number of components tested for these data was increased going from 1 to 5, thus enabling a more precise clustering of the population. The results, as shown in Table 2 , demonstrate strong reliability for the DPCM inference technique, achieving an average integrated Brier score of 0.0361, which is lower than its competitors.

** https://www.srtr.org/

5.3 | Interpretation of results

We now turn our attention to DPCM method (with $h = 2$ components). This model enables detailed analysis of the various variables in the data and the individuals grouped within the identified clusters.

While the number of components might be unexpected given the known heterogeneity of the population (one might have expected more groups) it may be explained by the relatively small number of patients included in the dataset, as we are working with only two years of transplantation data (representing 9,5% of the total number of patients in the database). Furthermore, the small representation of identified subgroups in the data, such as children—for whom we know the donor/recipient matching process cannot be handled in the same way as it is for adults[26] and who only represent 4.3% of the entire dataset population—may also have contributed to the selection of fewer components. The small number of children might be insufficient to have justified creating a distinct subgroup for them.

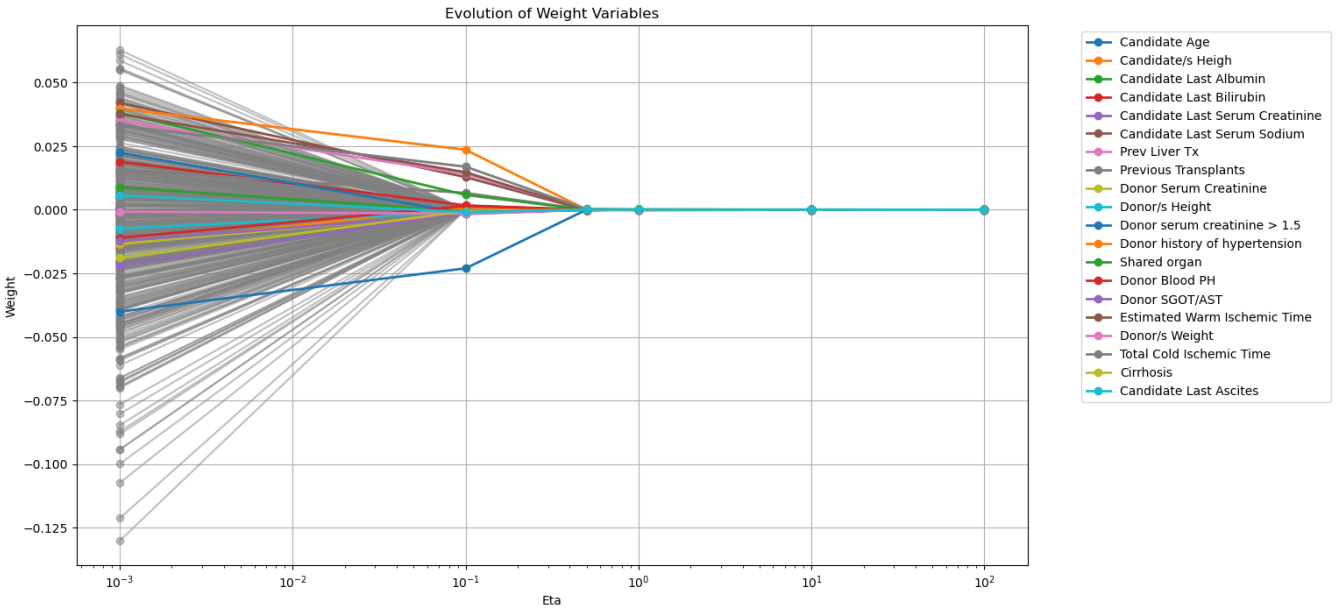

FIGURE 1 Evolution of weight of SRTR variables according to η

Analyzing the selected variables is then essential for the clinical field, as it helps in understanding the factors influencing the success and failure of graft transplantation. The graph presented in Figure 1 illustrates the evolution of variables weight in the model as the η values increase. This analysis is crucial because it not only deepens our understanding of why graft failures occur but also helps us identify key factors that contribute to both the success and failure of transplantation. Moreover, it confirms that the model behaves as expected by penalizing the weight of variables as η increases. These insights are vital for improving survival rates after transplantation, as they are not currently considered in the MELD score.

Notably, some of these variables, such as candidate's bilirubin creatinine or sodium, are identical to those used in calculating the MELD score, which is currently employed in clinical practice. Additionally, variables like SGOT (glutamic-oxaloacetic transaminase) provide information about the immunological status of patients, which is critical for donor-recipient matching during transplantation.

These findings validate the DPCM method by confirming existing knowledge within the medical community, indicating that the model's predictions are consistent with established practices. Furthermore, the study introduces novel elements by highlighting variables about ischemia time (warm and cold), which are linked to the preservation conditions during the transplantation process. These variables are not typically considered in current survival time evaluations but are known to be crucial during operations. Our study thus proposes a statistical model that incorporates both traditional factors, such as the MELD score, and new variables related to organ preservation, offering a more comprehensive approach to predicting patient survival.

6 | PERSPECTIVES

In this study, we investigated various Cox models, highlighting the significance of incorporating both population heterogeneity and penalization techniques. Accounting for heterogeneity in Cox models provides a deeper insight into survival data by identifying distinct subgroups with diverse risk profiles. This approach is essential for real-world scenarios, where patient populations are typically heterogeneous. By capturing and modeling these variations, we can achieve more precise and tailored survival predictions, enhancing the practical value of the models.

Additionally, integrating penalization techniques into the Cox framework presents notable benefits, particularly in handling high-dimensional data. Penalization helps reduce dimensionality, which is especially beneficial for identifying the most relevant variables in health-related datasets, where practitioners such as doctors or surgeons need to focus on a manageable set of predictors. Effective variable selection ensures that the models provide actionable insights without overwhelming clinicians with excessive information.

The combination of addressing heterogeneity and applying penalization represents a significant advancement in survival analysis. This dual strategy not only enhances predictive accuracy but also improves model interpretability and robustness. The flexibility of the DPCM method in adapting to different population structures and managing extensive predictor sets makes it particularly valuable for analyzing transplantation data. The findings of this study highlight the importance of these methodologies and support their continued development and application in understanding complex survival outcomes in the context of transplantation.

However, there are additional aspects that warrant further exploration. One area is the development of a penalization strategy that varies depending on the components of the mixture model, which could enhance model accuracy while maintaining interpretability. To achieve this, the penalization term should be made dependent on the components of the mixture model. This would allow for varying penalty strengths across different subgroups of the population, potentially leading to more precise results. However, this approach would significantly increase computational costs, as the optimization process would become more complex—requiring the determination of both the optimal number of components and the appropriate penalty factor for each. While this is a promising avenue for exploration it will be crucial to carefully plan how to manage the increased computational demands.

Additionally, while accounting for population heterogeneity is crucial for our model, the current method requires manual analysis of the identified subgroups to understand why each patient belongs to a specific cluster. This manual process is vital for predicting the survival time of new patients, as we must determine their appropriate cluster before applying the model. A significant advancement would be to develop a model that also predicts the subgroups for new patients as part of the inference process, rather than relying on subjective assignment. To achieve this, attention should focus on mixture of experts models[27], where the assignment of individuals to groups depends on their covariates. Such a model would simplify the prediction process and enhance the practical utility of our approach.

Another critical issue is the handling of missing data within the model, an aspect that is often encountered in clinical datasets. Addressing missing data is essential, as it can lead to biased estimates, reduced statistical power, and weakened model performance if not properly managed. In survival analysis, the naive approach of replacing missing values with a single imputation, such as the mean, can introduce bias, causing parameter estimates to shrink toward zero[28]. To overcome these issues, more advanced methods like multiple imputation[29] or techniques that incorporate the prediction of missing values during model inference should be utilized. Specifically, in our case, handling missing data within the EM algorithm by treating them as hidden variables to be estimated is an interesting approach. This method could significantly enhance the model's performance and would be especially useful in clinical settings where doctors may not have complete information about a patient but still need to predict survival time before transplantation.

In conclusion, the integration of mixture Cox models with L_1 penalization presents a powerful framework for analyzing complex survival data, particularly in the context of liver transplantation. While these advancements offer significant potential, addressing challenges such as component-specific penalization, mixture of experts, and the handling of missing data is crucial for further improving model accuracy and applicability. By refining these methods, we can enhance the predictive power and clinical relevance of survival models, ultimately contributing to better patient outcomes in transplantation and other medical fields.

DATA AVAILABILITY

The data that support the findings of this study were collected by the SRTR and were used with permission for this study. Data are not made available due to confidentiality restrictions.

ACKNOWLEDGMENTS

This study was supported by Institut Georges Lopez.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Trivedi HD. The evolution of the MELD score and its implications in liver transplant allocation: a beginner's guide for trainees. *ACG Case Reports Journal.* 2022;9(5):e00763.
2. Halldorson J, Bakthavatsalam R, Fix O, Reyes J, Perkins J. D-MELD, a simple predictor of post liver transplant mortality for optimization of donor/recipient matching. *American Journal of Transplantation.* 2009;9(2):318–326.
3. Flores A, Asrani SK. The donor risk index: a decade of experience. *Liver Transplantation.* 2017;23(9):1216–1225.
4. Solomon H. Opportunities and challenges of expanded criteria organs in liver and kidney transplantation as a response to organ shortage. *Missouri Medicine.* 2011;108(4):269–273.
5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association.* 1958;53(282):457–481.
6. Ishwaran H, Kogalur UB. Random survival forests for R. *R news.* 2007;7(2):25–31.
7. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.
8. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological).* 1972;34(2):187–202.
9. Cox DR. Partial Likelihood. *Biometrika.* 1975;62(2):269–276.
10. Therneau TM. Extending the Cox model. In: Springer. 1997:51–84.
11. McLachlan G. Finite mixture models. *A Wiley-Interscience Publication.* 2000.
12. Rosen O, Tanner M. Mixtures of proportional hazards regression models. *Statistics in Medicine.* 1999;18(9):1119–1131.
13. Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K. Deep Cox mixtures for survival regression. In: PMLR. 2021:674–708.
14. Ng SK, Xiang L, Yau KKW. *Mixture modelling for medical and health sciences*. Chapman and Hall/CRC, 2019.
15. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2. Springer, 2009.
16. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine.* 1997;16(4):385–395.
17. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal.* 2010;52(1):70–84.
18. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software.* 2011;39(5):1–20.
19. Webb A, Ma J, Lô SN. Penalized likelihood estimation of a mixture cure Cox model with partly interval censoring—An application to thin melanoma. *Statistics in Medicine.* 2022;41(17):3260–3280.
20. Collett D. *Modelling survival data in medical research*. CRC press, 2015.
21. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in medicine.* 2007;26(2):392–408.
22. Umgelter A, Hapfelmeier A, Kopp W, et al. Disparities in Eurotransplant liver transplantation wait-list outcome between patients with and without model for end-stage liver disease exceptions. *Liver Transplantation.* 2017;23(10):1256–1265.
23. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine.* 1999;18(17-18):2529–2545.
24. Squires JE, Bilhartz J, Soltys K, et al. Factors associated with improved patient and graft survival beyond 1 year in pediatric liver transplantation. *Liver Transplantation.* 2022;28(12):1899–1910.
25. Ravaioli M, Germinario G, Dajti G, et al. Hypothermic oxygenated perfusion in extended criteria donor liver transplantation—a randomized clinical trial. *American Journal of Transplantation.* 2022;22(10):2401–2408.
26. Barshes NR, Lee TC, Udell IW, et al. The pediatric end-stage liver disease (PELD) model as a predictor of survival benefit and posttransplant survival in pediatric liver transplant recipients. *Liver transplantation.* 2006;12(3):475–480.
27. Raman S, Fuchs TJ, Wild PJ, Dahl E, Buhmann JM, Roth V. Infinite mixture-of-experts model for sparse survival regression with application to breast cancer. *BMC bioinformatics.* 2010;11:1–10.
28. Ali A, Dawson S, Blows F, et al. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. *British journal of cancer.* 2011;104(4):693–699.
29. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in medicine.* 2009;28(15):1982–1998.
30. Breslow NE. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review.* 1975;43(1):45–57.
31. Städler N, Bühlmann P, Van De Geer S. 1-penalization for mixture regression models. *Test.* 2010;19:209–256.
32. Knaus WA, Harrell FE, Lynn J, et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine.* 1995;122(3):191–203.
33. Fleming TR, Harrington DP. *Counting processes and survival analysis*. 625. John Wiley & Sons, 2013.

How to cite this article: Peyraud E., Jacques J, Metzler G, Faivre I, and Dousse M. Mixture of Cox regression models with L_1-penalization for modeling patients survival time after liver transplantation. *Statistics in Medicine.* 2024.

APPENDIX

A COMPARAISON OF RUNNING TIMES

This appendix presents the running times for each of the models discussed in the main paper.

	PBC	**Framingham**	**SUPPORT**	**SRTR**
CPH (Linear)	1m 29s	10m 28s	4m 53s	26s
CPH (Non-Linear)	3m 1s	27m 3s	12m 34s	N/A
Penalized CPH (Linear)	9m 31s	52m 27s	31m 40s	25m 34s
Penalized CPH (Non-Linear)	17m 38s	2h 2m 6s	1h 12m 37s	N/A
Mixture CPH (Linear)	3m 7s	24m 13s	14m 42s	30m 26s
Mixture CPH (Non-Linear)	6m 9s	36m 6s	41m 7s	N/A
DPCM (Linear)	21m 36s	2h 28m 35s	1h 44m 17s	208m 14s
DPCM (Non-Linear)	1h 1m 40s	4h 17m 7s	4h 24m 26s	N/A

TABLE A1 Running times for each datasets (PBC, Framingham, SUPPORT, SRTR): Trained 10 times on linear and non-linear baselines for Cox Proportional Hazards Model (CPH), Mixtures of Cox Proportional Hazards Model (Mixture CPH), Penalized Cox Proportional Hazards Model (Penalized CPH), and Deep Penalized Cox Mixtures model (DPCM).

From details in Table A1 models incorporating penalization or mixture components consistently show longer training times compared to their basic counterparts. For example, Penalized Mixture CPH models generally require the most time to train, due to the increased computational complexity associated with both penalization and the mixture components.

Non-linear models tend to have longer running times than their linear counterparts. This is evident across all datasets, where non-linear versions of both basic and penalized models take significantly more time to compute.

For smaller datasets like PBC, the difference in running times between linear and non-linear models is notable, with non-linear models often taking approximately twice as long. For larger datasets such as SUPPORT, the time differences are more pronounced, with non-linear and penalized models extending training times to several hours.

The analysis of running times highlights a consistent trend across all datasets: penalized models and non-linear configurations generally require more time to train compared to their non-penalized and linear counterparts. This can be largely explained by the extensive range of penalty terms tested ($\eta = \{0, 0.001, 0.1, 0.5, 1, 10, 100\}$), which introduces additional complexity into the training process. Furthermore, the limited number of mixture components tested ($k = \{2, 3\}$) may have restricted the models' potential complexity compared to configurations with higher numbers of components.

The faster times observed for the SRTR database can be attributed to the use of only a portion of the complete dataset, suggesting that training times would be substantially longer if the entire database were utilized.