Biostatistics, 2026, **27**(1), kxag039
https://doi.org/10.1093/biostatistics/kxag039
Article

Penalized likelihood optimization for adaptive neighborhood clustering in time-to-event data with group-level heterogeneity

Alessandra Ragni[1,*], Lara Cavinato[1], and Francesca Ieva[1,2]

[1]MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy
[2]Health Data Science Centre, Human Technopole, Viale Rita Levi–Montalcini 1, Milan 20157, Italy

*Corresponding author: MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy. Email: alessandra.ragni@polimi.it.

Summary

The identification of patient subgroups with comparable event-risk dynamics plays a key role in supporting informed decision-making in clinical research. In such settings, it is important to account for the inherent dependence that arises when individuals are nested within higher-level units, such as hospitals. Existing survival models account for group-level heterogeneity through frailty terms but do not uncover latent patient subgroups, while most clustering methods ignore hierarchical structure and are not estimated jointly with survival outcomes. In this work, we introduce a new framework that simultaneously performs patient clustering and shared-frailty survival modeling through a penalized likelihood approach. The proposed methodology adaptively learns a patient-to-patient similarity matrix via a modified version of spectral clustering, enabling cluster formation directly from estimated risk profiles while accounting for group membership. A simulation study highlights the proposed model's ability to recover latent clusters and to correctly estimate hazard parameters. We apply our method to a large cohort of heart-failure patients hospitalized with COVID-19 between 2020 and 2021 in the Lombardy region (Italy), identifying clinically meaningful subgroups characterized by distinct risk profiles and highlighting the role of respiratory comorbidities and hospital-level variability in shaping mortality outcomes. This framework provides a flexible and interpretable tool for risk-based patient stratification in hierarchical data settings. Accompanying software is available at https://github.com/alessandragni/HazClust.

Keywords clustering, frailty models, heterogeneity, penalization, survival analysis

Received: January 12, 2026. **Revised**: August 26, 2026. **Accepted**: August 31, 2026
© The authors 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

1. INTRODUCTION

In real-world time-to-event datasets arising from clinical and epidemiological studies, identifying patients with similar risk trajectories enables the formation of subgroups, or clusters, characterized by shared likelihoods of experiencing key outcomes (eg death, relapse, treatment failure). Such stratification, relevant across a variety of application domains, in this setting can support clinical decision-making, facilitate disease subtyping, and advance precision-medicine initiatives (Collins and Varmus 2015). Nonetheless, patients are often intrinsically dependent, as they may share group membership—such as being treated in the same hospital—inducing correlations in outcomes and additional variability that must be accounted for in the analysis, as they can influence patients' risk profiles in complex ways.

This scenario is common in healthcare applications and is exemplified by the project motivating our work, *Enhance-Heart*, a study conducted in the Lombardy region of Northern Italy, which was severely affected during the early stages of the COVID-19 pandemic due to an uncontrolled spread of SARS-CoV-2 infections (Alteri et al. 2021). This study focuses on patients with heart failure (HF) and aims, among other objectives, to investigate the complex relationship between patients' clinical profiles and COVID-19-related mortality, while also accounting for differences across hospitals. Indeed, recent research has highlighted a significant interplay between heart failure and COVID-19 during the pandemic, with pre-existing HF complicating COVID-19 management and prognosis, and potentially leading to complications (Adeghate et al. 2021; Bader et al. 2021). Thus, identifying latent subgroups of HF patients hospitalized for COVID-19 while assessing how respiratory pathologies and overall healthcare condition influence mortality risk across patient clusters represents a key step in characterizing outcomes in a time-to-event framework. On the other hand, hospital-based resource availabilities, such as intensive care unit beds and general medicine/surgical beds, were found to be significantly associated with COVID-19 mortality incidence (Janke et al. 2021). Therefore, an accurate assessment of survival outcomes in this context requires modeling approaches capable of stratifying patient populations while accounting for differences across hospitals.

Traditional survival models often address heterogeneity arising from group-level dependence through a frailty term, that is a multiplicative factor in the hazard function, ie the risk of experiencing an event at a given time conditional on survival up to that point (Duchateau and Janssen 2008). This term can be specified in different ways, either parametrically—most commonly using Gamma or Lognormal distributions—or semiparametrically [see, eg Abrahantes et al. (2007); Balan and Putter (2020)]. However, these models often assume homogeneity within groups and overlook latent similar patterns between patients induced by their risk of experiencing an event of interest at a given time. Conversely, unsupervised stratification methods can reveal hidden patient subgroups by leveraging similarities across individuals; however, their limited integration with time-to-event outcomes often leads to suboptimal predictive performance. For example, Ahlqvist et al. (2018) performed a cluster analysis on 6 standardized clinical measures, identifying replicable diabetes subgroups with distinct prognoses and complication profiles. Advancing toward survival-informed clustering, Tosado et al. (2020) proposed transforming the feature space according to each feature's association with survival (via martingale residuals) before applying clustering in this transformed space. Similarly, Chen et al. (2016) introduced a recursive partitioning algorithm that derives prognostic cancer staging systems by maximizing survival separation between groups of patients. Shifting the focus toward prediction, Chapfuwa et al. (2020) developed a Bayesian nonparametric framework that embeds subjects in a clustered latent space to enable accurate time-to-event predictions and subgroups with distinct risk profiles. Along related lines, Manduchi et al. (2021) presented a variational-autoencoder-based survival clustering approach that jointly leverages patient features and time-to-event outcomes, while Buginga and de Souza e Silva (2024) proposed a nonparametric mixture-of-experts model in which each expert captures a distinct subgroup's survival distribution, thereby uncovering heterogeneous risk patterns. To the best of our knowledge, relatively few statistical models address the joint problem of clustering patients and modeling time-to-event outcomes while incorporating hierarchical structure. Among these,

the cluster-weighted model of Caldera et al. (2025) is a recent approach that fits a shared frailty model within each cluster and models covariates with cluster-specific parametric distributions, factorizing the joint density via cluster-specific mixing weights. While promising, its reliance on strong distributional assumptions for the covariates may restrict its flexibility.

To address this gap, we propose a novel model that extends parametric shared frailty models to account for group membership (eg patients treated in the same hospital) while simultaneously estimating both patient cluster membership—defined by similar risk trajectories—and hazard function parameters via penalized likelihood optimization, where the resulting solution is obtained through an iterative procedure. Patient stratification is achieved through a modified version of spectral clustering approach following the strategies proposed in Nie et al. (2014, 2016) and Guo (2015), consisting in the adaptive estimation of a similarity matrix capturing patient-to-patient relationships representing the probability that patients are *neighbors*, based on their estimated hazard risk, combining frailty and covariate effects. At the end of the procedure, graph-theoretic principles ensure that the connected components of the resulting graph, represented by the learned similarity matrix, correspond exactly to the number of clusters.

Our model builds upon the framework proposed by Liu et al. (2022), originally developed for cancer subtyping, which integrates supervised graph-based clustering into the modeling of multi-omic tumor data through penalization terms embedded in a *partial* likelihood-based optimization scheme. Related extensions and applications of this approach have been explored in radiomics settings in Cavinato et al. (2023). A key advantage of this framework is that patient clustering is embedded directly within the optimization procedure, so that cluster assignments are guided by underlying risk dynamics and reflect distinct event hazard profiles, making them interpretable in terms of event risk. Building on this framework, our contribution is 2-fold:

(i) We extend it to account for group-level (hospital) dependence, captured via a shared frailty term, together with patient-level heterogeneity (arising from subject-specific covariates and latent cluster membership) within a unified survival modeling framework;
(ii) We propose a penalized *full*-likelihood approach that embeds patient clustering directly within the hazard model, allowing similarity structures to be learned in a risk-driven manner and producing clusters that are interpretable in terms of event risk.

This paper is organized as follows. Section 2 outlines the proposed methodology, detailing the notation and the estimation procedure. In Section 3, we present a simulation study designed to evaluate the performance of the proposed methodology. In Section 4, we describe the *Enhance-Heart* administrative database that motivated our work and report the main results obtained with our method. Finally, Section 5 provides concluding remarks, discusses the strengths and limitations of the proposed approach, and outlines directions for future work.

2. METHODOLOGY

In this section, we outline the methodology: a recap on notation for parametric frailty models (Section 2.1), the model description (Section 2.2), the model optimization (Section 2.3) and details related to the model selection procedure (Section 2.4).

2.1. Notation

Let the observed data be $\mathcal{Z} = \{(y_{gi}, \delta_{gi}, \boldsymbol{x}_{gi}) \mid i = 1, \ldots, n_g,\ g = 1, \ldots, M\}$, where units $i = 1, \ldots, n_g$ are nested within groups $g = 1, \ldots, M$, and the total number of units is $N = \sum_{g=1}^{M} n_g$. For each unit, $y_{gi} = \min(t_{gi}, C^*_{gi})$ denotes the observed time to event or censoring, whichever comes first, and $\delta_{gi} = \mathbb{I}(t_{gi} \leq C^*_{gi})$ indicates whether the event was observed ($\delta_{gi} = 1$) or censored ($\delta_{gi} = 0$). The covariate vector $\boldsymbol{x}_{gi} \in \mathbb{R}^{d \times 1}$ may include both group-level and unit-specific components. We assume independent and noninformative censoring. To model the hazard rate of subject i in group g at time t, accounting for the presence of heterogeneity across groups, among the possible approaches

(eg, fixed effects, stratified), we consider the shared frailty model (Duchateau and Janssen 2008)

$$h_{gi}(t \mid u_g) = h_0(t;\psi)u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d\times 1}$ is the vector of fixed-effect coefficients, the u_g's, $g = 1,\dots,M$, are the actual values of a random variable with probability density $f_U(\cdot \mid \theta)$, that is the frailty distribution, and $h_0(t;\psi)$ is the baseline hazard, assumed parametric, depending on the vector of parameters ψ, and$'$ stands for the transpose. The cumulative hazard then follows as $H_{gi}(t \mid u_g) = u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})H_0(t;\psi)$, with $H_0(t;\psi) = \int_0^t h_0(s;\psi)ds$. Following computations in Section S1, the marginal loglikelihood can be expressed as

$$\begin{aligned}\ell(\psi,\boldsymbol{\beta},\theta \mid \mathcal{Z}) = \sum_{g=1}^{M}\Bigg[\Bigg[\sum_{i=1}^{n_g}\delta_{gi}\left(\log(h_0(y_{gi};\psi)) + \boldsymbol{x}'_{gi}\boldsymbol{\beta}\right)\Bigg] \\ + \log\left[(-1)^{d_g}\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)\right]\Bigg]\end{aligned} \tag{2.2}$$

where $d_g := \sum_{i=1}^{n_g}\delta_{gi}$ is the number of events in group g and $(-1)^r\mathcal{L}^{(r)}(s) := (-1)^r\frac{d^r}{ds^r}\mathcal{L}(s) = \mathbb{E}[U^r\exp(-Us)]$, being $\mathcal{L}(s)$ the Laplace transform, for $s \geq 0$, and r the order of derivative (Munda et al. 2012). Finally, it is known that the frailty term u_g can be predicted by $\hat{u}_g = \mathbb{E}\big(U \mid \mathcal{Z}_g;\hat{\psi},\hat{\boldsymbol{\beta}},\hat{\theta}\big)$, being $\mathcal{Z}_g$ the observed data in group g and

$$\mathbb{E}\big(U \mid \mathcal{Z}_g;\boldsymbol{\beta},\psi,\theta\big) = -\frac{\mathcal{L}^{(d_g+1)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)}{\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)}. \tag{2.3}$$

2.2. Model

In a parametric setting, such as that addressed by Munda et al. (2012), where both the frailty and the baseline hazard follow specified parametric distribution, the model parameters $\boldsymbol{\beta},\psi,\theta$ can be estimated by maximizing the loglikelihood in (2.2), or minimizing the negative loglikelihood: $\min_{\boldsymbol{\beta},\psi,\theta}\{-\ell(\psi,\boldsymbol{\beta},\theta \mid \mathcal{Z})\}$.

With the aim of performing clustering on the units while learning $\boldsymbol{\beta},\psi$ and θ, let $\mathbf{S} \in \mathbb{R}^{N\times N}$ be the affinity matrix whose entries $\{s_{gi,hj}\}$ represent the similarity between the unit i in group g and the unit j in group h. Specifically, we interpret each entry $\{s_{gi,hj}\}$ as the probability that unit i in group g has unit j in group h as a neighbor, imposing the constraint $\sum_{h=1}^{M}\sum_{j=1}^{n_h}s_{gi,hj} = 1$, $\forall g,i$ that ensures that the connection probabilities for each unit sum up to 1. By assuming that units have a higher probability of being neighbors if they are more similar, or equivalently have a lower distance $d_{gi,hj}(\boldsymbol{\beta},\theta,\mathcal{Z})$, we rewrite the above optimization problem by further penalizing the loglikelihood for jointly estimating $\boldsymbol{\beta},\psi,\theta$, and $\mathbf{S}$:

$$\begin{aligned}&\min_{\boldsymbol{\beta},\psi,\theta,\mathbf{S}} -\ell(\psi,\boldsymbol{\beta},\theta \mid \mathcal{Z}) + \gamma\sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left(d_{gi,hj}(\boldsymbol{\beta},\theta,\mathcal{Z})\,s_{gi,hj} + \mu\,s_{gi,hj}^2\right) \\ &\text{s.t.}\sum_{h=1}^{M}\sum_{j=1}^{n_h}s_{gi,hj} = 1, \quad \mathbf{s}_{gi} \geq 0, \quad \forall i = 1,\dots,n_g,\ g = 1,\dots,M\end{aligned} \tag{2.4}$$

where $\mathbf{s}_{gi} \in \mathbb{R}^{1\times N}$ is the row of $\mathbf{S}$ corresponding to unit i in group g, and the reported inequality is elementwise. The last 2 terms in first line of (2.4) follow the literature on clustering with adaptive neighbors, see for instance Nie et al. (2014). These terms regulate the structure of neighborhood probabilities in $\mathbf{S}$: the first term encourages higher probabilities for nearby units, that, if alone leads to the limit case of 1 for the nearest neighbor and 0 elsewhere, while the quadratic term promotes smoother and more evenly distributed assignments, with the limit case assigning equal probability

$1/N$ to all units. Since the entries of $\mathbf{S}$ are unknown neighborhood probabilities to be estimated, the distance term provides the information driving their learning. Including it explicitly in the objective function enables joint estimation of the similarity structure and the survival model parameters. The hyperparameter γ controls the strength of the penalization; when $\gamma = 0$, no penalization is applied, and consequently no clustering is induced, as the penalty and its associated constraints vanish. In contrast, μ is a tunable parameter that modulates the influence of the similarity structure in the overall optimization. Its selection will be discussed in Sections 2.3 and 2.4.

As we aim to learn $\mathbf{S}$ based on the similarity between the risk of experiencing the event of interest for patient i in group g, compared to patient j in group h, that we quantify on the log-hazard scale, as $\log h_{gi}(t \mid u_g) = \log h_0(t;\psi) + \log u_g + \boldsymbol{x}'_{gi}\boldsymbol{\beta}$ and $\log h_{hj}(t \mid u_h)$, where the $\log h_0(t;\psi)$ shared across units cancels out in pairwise comparisons. Thus, we define

$$d_{gi,hj}(\boldsymbol{\beta},\theta,\mathcal{Z}) := \left\| (\boldsymbol{x}'_{gi}\boldsymbol{\beta} + \log u_g) - (\boldsymbol{x}'_{hj}\boldsymbol{\beta} + \log u_h) \right\|^2, \tag{2.5}$$

where $\|\cdot\|$ denotes the Euclidean distance, and u_g and u_h are the frailty terms defined in (2.3). These terms ensure that the distance captures between-group hazard differences: they cancel within a group but persist across groups, thereby encoding unobserved heterogeneity and promoting a similarity graph that aligns with frailty-adjusted risk profiles.

To perform clustering into C clusters, we interpret $\mathbf{S} \in \mathbb{R}^{N\times N}$ as the adjacency matrix of a graph with the N units as nodes. In this setting, we define the Laplacian matrix $\mathbf{L_S} := \mathbf{D_S} - (\mathbf{S}' + \mathbf{S})/2$, where $\mathbf{D_S} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix, where the gi-th entry is $\sum_{hj}(s_{gi,hj} + s_{hj,gi})/2$. Following Mohar et al. (1991), the multiplicity C of the zero eigenvalue of $\mathbf{L_S}$ corresponds to the number of connected components in the graph with similarity matrix $\mathbf{S}$. Enforcing $rank(\mathbf{L_S}) = N - C$ thus promotes a partition into C clusters. Intuitively, clustering corresponds to partitioning the similarity graph defined by $\mathbf{S}$ into groups of patients with comparable *frailty-adjusted risk profiles*, with membership given by the C connected components induced by the Laplacian rank constraint, as in *spectral clustering*. Furthermore, as shown in Nie et al. (2014) and Fan (1949), this rank constraint can be handled equivalently and more easily by introducing a matrix $\mathbf{F} \in \mathbb{R}^{N\times C}$, given that

$$2Tr(\mathbf{F}'\mathbf{L_S}\mathbf{F}) = \sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h} \|\mathbf{f}_{gi} - \mathbf{f}_{hj}\|^2 s_{gi,hj}, \tag{2.6}$$

being $\mathbf{f}_{gi} \in \mathbb{R}^{1\times C}$ the row of $\mathbf{F}$ corresponding to the belonging of unit i in group g.

Combining (2.2), (2.4), and (2.5), we get an optimization problem that reads

$$\begin{aligned}
\min_{\boldsymbol{\beta},\psi,\theta,\mathbf{S},\mathbf{F}} & -\left\{\sum_{g=1}^{M}\left[\left[\sum_{i=1}^{n_g}\delta_{gi}\left(\log(h_0(y_{gi};\psi)) + \boldsymbol{x}'_{gi}\boldsymbol{\beta}\right)\right] + \log\left[(-1)^{d_g}\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g} H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)\right]\right]\right\} \\
& + \gamma\left\{\sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[\left\|(\boldsymbol{x}'_{gi}\boldsymbol{\beta} + \log u_g) - (\boldsymbol{x}'_{hj}\boldsymbol{\beta} + \log u_h)\right\|^2 s_{gi,hj} + \mu\, s^2_{gi,hj}\right] + 2\lambda\, Tr(\mathbf{F}'\mathbf{L_S}\mathbf{F})\right\} \\
\text{s.t.} & \sum_{h=1}^{M}\sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad \mathbf{s}_{gi} \geq 0, \quad \forall i = 1,\ldots,n_g,\ g = 1,\ldots,M, \quad \mathbf{F}'\mathbf{F} = \mathbf{I},\ \mathbf{F} \in \mathbb{R}^{N\times C}
\end{aligned} \tag{2.7}$$

Following from (2.3), and noting that $-1 = \frac{(-1)^{d_g+1}}{(-1)^{d_g}}$, we apply a logarithmic property to obtain

$$\begin{aligned}
\log u_g = & \log\left[(-1)^{d_g+1}\mathcal{L}^{(d_g+1)}\left(\sum_{i=1}^{n_g} H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)\right] \\
& - \log\left[(-1)^{d_g}\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g} H_0(y_{gi};\psi)\exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)\right].
\end{aligned}$$

2.3. Optimization procedure

To solve the optimization problem in (2.7), we adopt a block coordinate descent approach: at each iteration, we update 1 block of variables while keeping the others fixed, until convergence. The procedure alternates among the following 3 steps:

Step 1 Fix $\mathbf{S}, \beta, \psi, \theta$ and update $\mathbf{F}$. With $\mathbf{S}$, β, ψ, and θ fixed, the update of $\mathbf{F}$ reduces to the following spectral optimization problem (Nie et al. 2014):

$$\min_{\mathbf{F}} \; 2\lambda Tr(\mathbf{F}^T \mathbf{L}_{\mathbf{S}} \mathbf{F}) \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{F} \in \mathbb{R}^{N\times C}$$

The factor 2λ does not affect the solution and can be omitted. The optimal $\mathbf{F}$ is obtained by taking the C eigenvectors of the Laplacian $\mathbf{L}_{\mathbf{S}}$ associated with its C smallest eigenvalues. This step requires the number of clusters C and the current estimate of $\mathbf{S}$.

Step 2 Fix $\mathbf{F}, \mathbf{S}$ and update β, ψ, θ. With $\mathbf{F}$ and $\mathbf{S}$ fixed, the optimization over (β, ψ, θ) reduces to a survival modeling problem with an additional similarity-based regularization term. The optimization problem can be expressed as:

$$\min_{\beta,\psi,\theta} \; -\ell(\psi, \beta, \theta \mid \mathcal{Z}) + \gamma \sum_{g,h=1}^{M} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \left(\left\| \boldsymbol{x}'_{gi}\beta + \log u_g - \boldsymbol{x}'_{hj}\beta - \log u_h \right\|^2 \right) s_{gi,hj}$$

We estimate the survival model parameters relying on maximum likelihood optimization with parametric frailty, following Munda et al. (2012), and extend this framework with the similarity-based penalty. This step requires γ, the current similarity matrix $\mathbf{S}$, the dataset and model specification.

Step 3 Fix $\mathbf{F}, \beta, \psi, \theta$ and update $\mathbf{S}$. With $\mathbf{F}$, β, ψ, and θ fixed, the update of $\mathbf{S}$ reads:

$$\min_{\mathbf{S}} \; \gamma \left\{ \sum_{g,h=1}^{M} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \left[\left\| \boldsymbol{x}'_{gi}\beta + \log u_g - \boldsymbol{x}'_{hj}\beta - \log u_h \right\|^2 s_{gi,hj} + \mu s^2_{gi,hj} \right] + 2\lambda Tr(\mathbf{F}'\mathbf{L}_{\mathbf{S}}\mathbf{F}) \right\}$$

$$\text{s.t.} \sum_{h=1}^{M} \sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad 0 \le \mathbf{s}_{gi} \le 1, \quad \forall i = 1, \ldots, n_g, \; g = 1, \ldots, M$$

Using relation (2.6) and the separability of the problem across units, this reduces to an independent optimization for each $\mathbf{s}_{gi}$:

$$\min_{\mathbf{s}_{gi}} \sum_{h=1}^{M} \sum_{j=1}^{n_h} \left(\left\| \boldsymbol{x}'_{gi}\beta + \log u_g - \boldsymbol{x}'_{hj}\beta - \log u_h \right\|^2 + \mu s_{gi,hj} + \lambda \|\mathbf{f}_{gi} - \mathbf{f}_{hj}\|^2 \right) s_{gi,hj}$$

$$\text{s.t.} \sum_{h=1}^{M} \sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad \mathbf{s}_{gi} \ge 0 \tag{2.8}$$

Let $w_{gi,hj}$ be the sum of the first and third terms inside the parentheses of the objective in (2.8) and $\boldsymbol{w}_{gi} \in \mathbb{R}^{N\times 1}$. As shown in Section S2, problem (2.8) admits a closed-form solution for each row $\mathbf{s}_{gi}$: $s_{gi,hj} = \max\{0, \alpha_{gi} - w_{gi,hj}/(2\mu_{gi})\}$ subject to $\mathbf{s}'_{gi}\mathbf{1} = 1$, where $\alpha_{gi} \ge 0$ and μ_{gi} may differ across units.

By enforcing k-nearest-neighbor sparsity on $\mathbf{s}_{gi}$, ie allowing only the k nearest neighbors to connect unit i in group g, following Nie et al. (2014) and Guo (2015), letting $\hat{\boldsymbol{w}}_{gi} = (\hat{w}_{gi,1}, \ldots, \hat{w}_{gi,N-1})$ denote the sorted distances (excluding the self-distance), and imposing $\mathbf{s}'_{gi}\mathbf{1} = 1$, as shown in Section S3, we obtain

$$\alpha_{gi} = \frac{1}{k} + \frac{1}{2k\mu_{gi}} \sum_{p=1}^{k} \hat{w}_{gi,p} \quad \text{and} \quad \mu_{gi} = \frac{k}{2}\hat{w}_{gi,(k+1)} - \frac{1}{2}\sum_{p=1}^{k} \hat{w}_{gi,p}. \tag{2.9}$$

Finally, once all rows are updated, we set $\mu = \sum_{g,i} \mu_{gi}/N$. This step requires λ, k, the current estimate of $\mathbf{F}$, the dataset and the model specification. Practical guidance on how to choose k is provided in Section S4.

Convergence is assessed based on both the stability of the similarity matrix and the quality of the statistical fit. Specifically, the algorithm terminates when all of the following conditions are satisfied: (i) the change in $\mathbf{S}$ between successive iterations is negligible; (ii) the improvement in log-likelihood is below a prescribed threshold; and (iii) the number of connected components in $\mathbf{S}$ equals the target C. If these conditions are not met, the procedure stops after a maximum number of iterations. Further computational details, including implementation settings and inference procedures, are provided in Section S5.

2.4. Model selection

To assess clustering quality across different model configurations (parametric distributions, γ, k, and C), we use the silhouette coefficient (Kaufman and Rousseeuw 2009), a widely used measure balancing within-cluster cohesion and between-cluster separation. Given a unit i, let $a(i)$ denote its average distance to all other units in the same cluster, and let $b(i)$ be the minimum average distance from i to any other cluster. We recall that the silhouette value is defined as $s(i) = [b(i) - a(i)]/\max\{a(i), b(i)\}$, $s(i) \in [-1, 1]$, with higher values indicating better separation. In our setting, distances for unit i with respect to units j are computed as $\sqrt{d_{gi,hj}(\boldsymbol{\beta}, \theta, \mathcal{Z})}$ in (2.5), as it incorporates the covariate effects and the estimated frailty contributions. Group indices (such as g for i and h for j) are omitted here since they play no role in the silhouette calculation. We then assess the overall clustering quality using the mean silhouette, and select the optimal clustering configuration as the one with the maximum mean silhouette, thereby achieving the best separation with respect to the fitted survival model.

3. SIMULATION STUDY

We here present a simulation study to assess the performance of the proposed methodology.

Aims. Our main objectives are to evaluate, under varying penalization strengths (γ), the ability of the method (i) to correctly identify the underlying cluster structure, including the true number of clusters, and (ii) to accurately estimate model unknown parameters $(\boldsymbol{\beta}, \boldsymbol{\psi}, \theta)$.

Data-generating mechanism. We generate data by inducing the presence of $M = 10$ groups, each containing $n_g = 50$ ($g = 1, \ldots, M$), yielding a total sample size of $N = 500$. To induce dependence among units within the same group, we introduce a group-level frailty u_g, drawn independently for each group from a Gamma distribution $u_g \sim \text{Gamma}(1/\theta, 1/\theta)$, with $\theta = 0.5$ so that $\mathbb{E}[u_g] = 1$ and $\text{Var}(u_g) = \theta$, inducing a moderate degree of within-group dependence.

Within each group g, every unit i is randomly assigned to one of the $C = 3$ latent clusters, denoted $z_{gi} \in \{1, 2, 3\}$. Each unit is associated with a covariate vector $\boldsymbol{x}_{gi}$ and a linear predictor $\eta_{gi} = \boldsymbol{x}'_{gi}\boldsymbol{\beta} + \log u_g$ where $\boldsymbol{\beta}$ is a vector of regression coefficients, so that, conditional on the cluster assignment z_{gi} and frailty u_g, the linear predictor is modeled as $\eta_{gi} \mid z_{gi} = c, u_g \sim \mathcal{N}(\mu_c, \sigma^2)$, $c = 1, 2, 3$ where $[\mu_1, \mu_2, \mu_3] = [-8, 0, 8]$ and $\sigma = 1$. This specification generates clearly separated distributions of η_{gi} across clusters, maximizing the clustering structure as measured in (2.5); additional scenarios, including weaker cluster separation, increased frailty variance, class imbalance, and model misspecification, are presented in Section S6.1. Finally, to preserve the intended simulation structure across clusters, we construct a single observed covariate x_{gi} as $x_{gi} = (\eta_{gi} - \log u_g)/\beta$, setting $\beta = \log(2)$.

Survival times are generated under a Weibull baseline hazard, parameterized by $\boldsymbol{\psi} = [\rho, \xi]$ through $H_0(t) = \xi^\rho t^\rho$ ($\rho, \xi > 0$), $\rho = 2.5$ and $\xi = 0.01$. The corresponding inverse cumulative hazard is $H_0^{-1}(s) = s^{1/\rho}/\xi$. Event times are generated using inverse transform sampling: we draw $V_{gi} \sim \text{Uniform}(0, 1)$ and set $t_{gi} = H_0^{-1}\big(-\log(V_{gi})/\exp(\eta_{gi})\big)$.

Censoring is imposed using 1 of 2 possible mechanisms: (i) *administrative* censoring at the fixed time of 100, and (ii) *Gaussian (normal)* censoring with mean 130 and standard deviation 15, analogous to defaults of `genfrail` in `frailtySurv` `R` package. Under this data-generating

mechanism, the Kaplan–Meier curves highlight distinct patterns across the 3 clusters: one maintains high survival, another declines gradually, and a third drops sharply early on.

Estimands and performance measures. The primary targets of estimation are the cluster assignments and the vector of model parameters. Performance is assessed through:

- *Cluster identification*, evaluated by the frequency with which convergence to the true number of clusters ($C = 3$) is reached, and by the mean silhouette width (see Section 2.4).
- *Clustering quality*, assessed conditionally on correct identification of $C = 3$, using the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and accuracy. The ARI, computed via scikit-learn, measures pairwise agreement between estimated and true partitions, adjusting for chance; it ranges from -0.5 (worse than chance) to 1 (perfect agreement), with 0 indicating random labeling. Accuracy measures the proportion of correctly classified observations.
- *Parameter estimation*, evaluated in terms of bias and variability of the estimated parameters, and summarized through boxplots.

Methods and implementation. We use a grid search over $C \in \{2, 3, 4, 5\}$, $k = \{20, 50\}$ and $\gamma \in \{0, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.4\}$, generating $B = 100$ independent datasets for each configuration. The case $\gamma = 0$ corresponds to absence of penalization, so no clustering is performed, and just a parametric frailty model is fitted; in this case, C and k play no role in the estimation. Convergence is assessed using standard stopping criteria ($\texttt{tolS} = 10^{-2}$, $\texttt{tolll} = 1$, $\texttt{maxit} = 200$, see Section S5).

3.1. Results

3.1.1. Cluster identification

We primarily evaluate the ability of the method to correctly identify the underlying cluster structure, including the true number of clusters.

We first focus on the simulations obtained by superimposing $C = 3$ and $\gamma > 0$. For both $k = 20$ and $k = 50$, the method consistently identifies 3 clusters, with only minor differences in the γ thresholds at which performance begins to decline. Under *administrative* censoring, the method identifies the 3 clusters with 100% frequency for $\gamma \leq 0.1$. For larger γ, the identification rate decreases slightly, reaching a minimum of 95% at $\gamma = 0.4$. Overall, convergence was achieved after an average of 17.49 iterations. Under *normal* censoring, perfect identification is maintained for $\gamma \leq 10^{-3}$, after which the rate gradually slightly declines, reaching a minimum of 98% at $\gamma = 0.4$. In this second case, the algorithm required on average 24.49 iterations to converge.

We further assess cluster identification ability using the mean silhouette width under varying choices of C, fixing $k = 20$ and $\gamma = 10^{-4}$. The results, reported in Table 1, indicate that $C = 3$ yields the highest mean and median silhouette values under both censoring mechanisms, suggesting the best overall separation of clusters. Although the silhouette values for $C = 3$ exhibit slightly higher variability, the overall separation remains substantially stronger. Moreover, convergence to the selected number of clusters is always achieved for $C = 3$, while this is not uniformly the case for alternative values of C.

3.1.2. Clustering quality

We now evaluate clustering quality conditional on correct identification of $C = 3$, using accuracy and the ARI. Results are reported in Table 2 across different censoring mechanisms, values of k and γ. Clustering performance is generally very high, especially for larger k, with accuracy and ARI frequently approaching 1. Specifically, the median values remain close to 1 across most settings, reflecting the skewed distribution of performance measures near the upper bound; the mean values, in contrast, show the gradual rise and the sharp drop more clearly: performance is slightly lower for very small γ, improves for intermediate values, and decreases sharply at $\gamma = 0.4$, where penalization becomes much stronger. Standard deviations are very small when performance is high, indicating stable behavior across replications.

Additional simulation settings are reported in Section S6.1. Overall, results suggest that the proposed approach is robust across the scenarios considered, and consistently outperforms a

Table 1 Summary of the empirical distribution of the mean silhouette values (range in $[-1, 1]$) over $B = 100$ replications under 2 censoring mechanisms—(i) administrative and (ii) normal—for $k = 20$ and $\gamma = 10^{-4}$. For each value of C, the table reports the mean, median, standard deviation, and the proportion of runs that converged to the selected C within maxit iterations. The estimates related to $C = 3$ are highlighted in bold.

Censoring	C	Mean silhouette			Convergence
		Mean	Median	SD	%
(i) Administrative	2	0.660	0.670	0.046	67
	3	**0.792**	**0.832**	**0.116**	**100**
	4	0.661	0.718	0.121	100
	5	0.567	0.612	0.113	100
(ii) Normal	2	0.618	0.649	0.082	93
	3	**0.670**	**0.789**	**0.182**	**100**
	4	0.541	0.577	0.173	99
	5	0.466	0.484	0.145	100

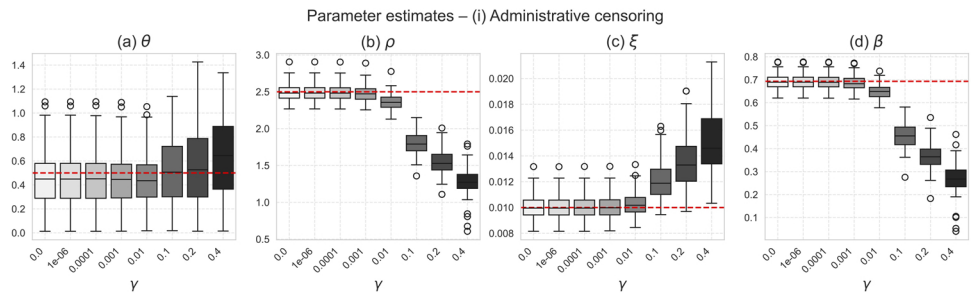


Figure 1 Boxplots of parameter estimates for (a) θ, (b) ρ, (c) ξ, and (d) β across penalization levels γ (x-axis) under *administrative* censoring for $k = 20$ and $C = 3$. The true parameter values are shown as a piecewise-constant horizontal line.

benchmark based on a 2-stage procedure, with performance declining mainly when the clustering signal is weak or the class balance is severely distorted.

3.1.3. Parameter estimation

Finally, we evaluate parameter estimation accuracy focusing on the case $C = 3$, and restricting the attention to $k = 20$ to ease visualization and reduce the number of scenarios presented. We examine the estimates of the unknown parameters in the shared frailty model, $[\theta, \rho, \xi, \beta]$ (with $\psi = [\rho, \xi]$); for completeness, we also include the unpenalized case $\gamma = 0$, corresponding to a standard parametric frailty model. Figures 1 and S6.1 (in Section S6.2) report boxplots of the estimated parameters across different values of γ under administrative and normal censoring, respectively. The true parameter values are indicated by horizontal dashed lines. As expected, the bias generally increases with γ. In particular, θ exhibits slight underestimation even in the unpenalized case, consistent with previous findings based on the `parfm` implementation (Munda et al. 2012). Most parameters show stable variability across penalization levels, with the exception of ξ, which becomes increasingly variable as γ increases. To further quantify the effects of penalization on estimation accuracy, Section S6.3 analyzes the variance and mean squared error-to-variance ratio across all γ values and censoring schemes.

4. CASE STUDY

This section applies the proposed methodology to the *Enhance-Heart* dataset, a large administrative dataset linked with the COVID-19 registry (https://www.dati.lombardia.it/) of Lombardy Region (LR in the following).

Table 2 Empirical means, medians, and standard deviations of *accuracy* and *ARI* over $B = 100$ replications, evaluated under 2 censoring mechanisms—(i) administrative and (ii) normal—and across varying values of k and γ, setting $C = 3$. The best estimates in each setting are highlighted in bold.

Censoring	k	γ	Accuracy			ARI		
			Mean	**Median**	**SD**	**Mean**	**Median**	**SD**
(i) Administrative	20	10^{-6}	0.971	1.000	0.089	0.942	1.000	0.161
		10^{-4}	0.971	1.000	0.089	0.942	1.000	0.161
		10^{-3}	**0.975**	**1.000**	**0.081**	**0.950**	**1.000**	**0.147**
		0.01	0.957	1.000	0.082	0.908	1.000	0.172
		0.1	0.950	1.000	0.083	0.888	1.000	0.179
		0.2	0.930	1.000	0.115	0.860	1.000	0.197
		0.4	0.891	0.928	0.144	0.788	0.819	0.242
	50	10^{-6}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		10^{-4}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		10^{-3}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		0.01	1.000	1.000	0.001	0.999	1.000	0.004
		0.1	0.996	1.000	0.023	0.989	1.000	0.054
		0.2	0.945	0.998	0.120	0.897	0.994	0.199
		0.4	0.420	0.390	0.109	0.060	0.019	0.159
(ii) Normal	20	10^{-6}	0.902	0.998	0.142	0.811	0.994	0.246
		10^{-4}	0.904	0.998	0.138	0.810	0.994	0.251
		10^{-3}	0.891	0.970	0.141	0.788	0.918	0.243
		0.01	0.896	0.998	0.149	0.807	0.994	0.239
		0.1	0.913	**1.000**	0.117	0.820	**1.000**	0.224
		0.2	**0.922**	0.998	**0.097**	**0.831**	0.994	**0.195**
		0.4	0.870	0.894	0.145	0.752	0.732	0.237
	50	10^{-6}	0.990	1.000	0.040	0.976	1.000	0.094
		10^{-4}	0.990	1.000	0.040	0.977	1.000	0.094
		10^{-3}	0.990	1.000	0.041	0.976	1.000	0.096
		0.01	**0.998**	**1.000**	**0.016**	**0.994**	**1.000**	**0.034**
		0.1	0.994	1.000	0.027	0.986	1.000	0.060
		0.2	0.964	0.998	0.089	0.922	0.994	0.183
		0.4	0.417	0.386	0.102	0.056	0.015	0.148

4.1. Cohort selection

Sourcing from the administrative dataset related to the *Enhance-Heart* project, we select those patients that between *1 January 2018*, and *31 December 2020* were diagnosed with HF, that is, hospitalized or visited emergency room under DRG code 127 ("Heart Failure and Shock") according to the LR MS-DRG v40 system, thus including both primary/secondary diagnoses of HF (ICD-9-CM: 428.xx) and related conditions (ICD-9-CM: 402.01, 402.11, 402.91). Among these, we retain those hospitalized for COVID-19 according to a binary flag provided in the database, between *31 January 2020* (start of the national COVID-19 state of emergency in Italy) and *18 June 2021* (implementation of the Green Pass regulations governing reopening). For those patients, outcomes are assessed over a 90-d follow-up period; death within this window is considered the event, while patients surviving beyond are regarded as censored (55.25% of the sample).

The analysis is restricted to hospitals with a minimum of 50 patients hospitalized, as smaller hospitals may exhibit different organizational conditions and characteristics, and to ensure more robust results for the frailty-related estimates. We further exclude patients who required hospital

Table 3 Characteristics of the identified patient clusters with respect to variables described in Section 4.1 and summarized in Table S7.1. Continuous variables are summarized as mean (SD) and categorical variables as percentage. Statistical differences among clusters were assessed using one-way ANOVA, Kruskal–Wallis, or chi-square tests, depending on the variable type and distribution. Reported *P*-values correspond to the comparison across all 3 clusters.

	Cluster 1	**Cluster 2**	**Cluster 3**	***P*-value**
Number of patients	1077 (34.89%)	1054 (34.15%)	955 (30.94%)	...
Age at admission in years	83.5 (7.7)	81.8 (7.8)	78.4 (8.6)	<0.001
Chronic disease burden score (ModMCS)	12.6 (8.5)	11.8 (7.7)	9.8 (6.7)	<0.001
Respiratory disease presence (Resp)	0.99 (1.17)	1.03 (1.21)	0.87 (1.14)	0.005
Survival/observation time	0.15 (0.10)	0.16 (0.10)	0.18 (0.10)	<0.001
Male sex (Gender = male)	62.0%	57.4%	47.1%	<0.001
All-cause mortality (Status = 1)	51.4%	45.3%	36.6%	<0.001

transfers for medical purposes or who had multiple episodes of infection. The final cohort includes $N = 3086$ HF patients hospitalized for COVID-19 in $M = 32$ different hospitals in the LR.

The time-to-event outcome is defined by the pair (`Time`, `Status`), where `Time` denotes days until all-cause death within the 90-d observation window or censoring, and `Status` indicates event occurrence through a binary variable, equal to 1 if death from any cause occurs within the observation period, 0 otherwise. Patient's covariates include `Age` at admission, `Gender`, and Modified Multisource Comorbidity Score (`ModMCS`) (Corrao et al. 2017; Caldera et al. 2025), a measure of overall comorbidity burden unrelated to the primary diagnosis and a proxy for general health status, with higher values reflecting greater nonrespiratory comorbidity burden. To complement this, we captured specific respiratory conditions separately using a binary indicator, `Resp`, which equals 1 if a patient presents with more than one of the following conditions: chronic obstructive pulmonary disease, bronchitis, pneumonia, or respiratory failure. See Table S7.1 for detailed variable descriptions and summary statistics.

4.2. Model setup

We fit our model under multiple configurations and identify the optimal one as that with the maximum mean silhouette value, as described in Section 2.4 (Kaufman and Rousseeuw 2009). Specifically, we consider 2 choices for the frailty distribution (Gamma and Inverse Gaussian) and 2 options for the baseline hazard distribution (Weibull and Exponential). We evaluate models with $\gamma = \{10^{-8}, 10^{-5}, 10^{-2}, 10^{-1}\}$, $k \in \{50, 150\}$, and $C \in \{2, 3, 4\}$. We set `maxit` = 200, `tolS` = 10^{-2}, `tolll` = 1 and leave all the other parameters with default settings. The mean silhouette values for all the above configurations are reported in Table S7.2.

We select the model with the frailty distributed as Inverse Gaussian, the baseline as Weibull, $\gamma = 10^{-1}$, $k = 150$ and $C = 3$ as the optimal clustering configuration, as it yields the highest mean silhouette score equal to 0.609 and therefore the best separation.

4.3. Results

With the chosen model, we obtain the following regression coefficients: $\hat{\beta}_{\texttt{Age}} = 0.0025$, $\hat{\beta}_{\texttt{Gender=male}} = 0.0244$, $\hat{\beta}_{\texttt{ModMCS}} = 0.0013$, and $\hat{\beta}_{\texttt{Resp}} = 0.0033$. The corresponding *P*-values are all < 0.05, indicating statistically significant effects. The estimated frailty variance is $\theta = 0.573$, while the Weibull parameters were $\rho = 0.618$ and $\lambda := \xi^{\rho} = 1.126$, respectively, following parametrization in Tables S5.1 and S5.2. The model reaches convergence in 23 iterations.

The resulting clusters—corresponding to disconnected components of the learned patient similarity graph—contain 1077, 1054, and 955 patients respectively. Cluster 1, the largest, comprises the oldest patients with the highest comorbidity burden and shortest mean follow-up; Cluster 2 includes slightly younger patients with the highest respiratory burden; Cluster 3 comprises the youngest patients with the lowest comorbidity and respiratory burden and the longest mean follow-up. All between-cluster differences are statistically significant ($P < 0.05$; Table 3).

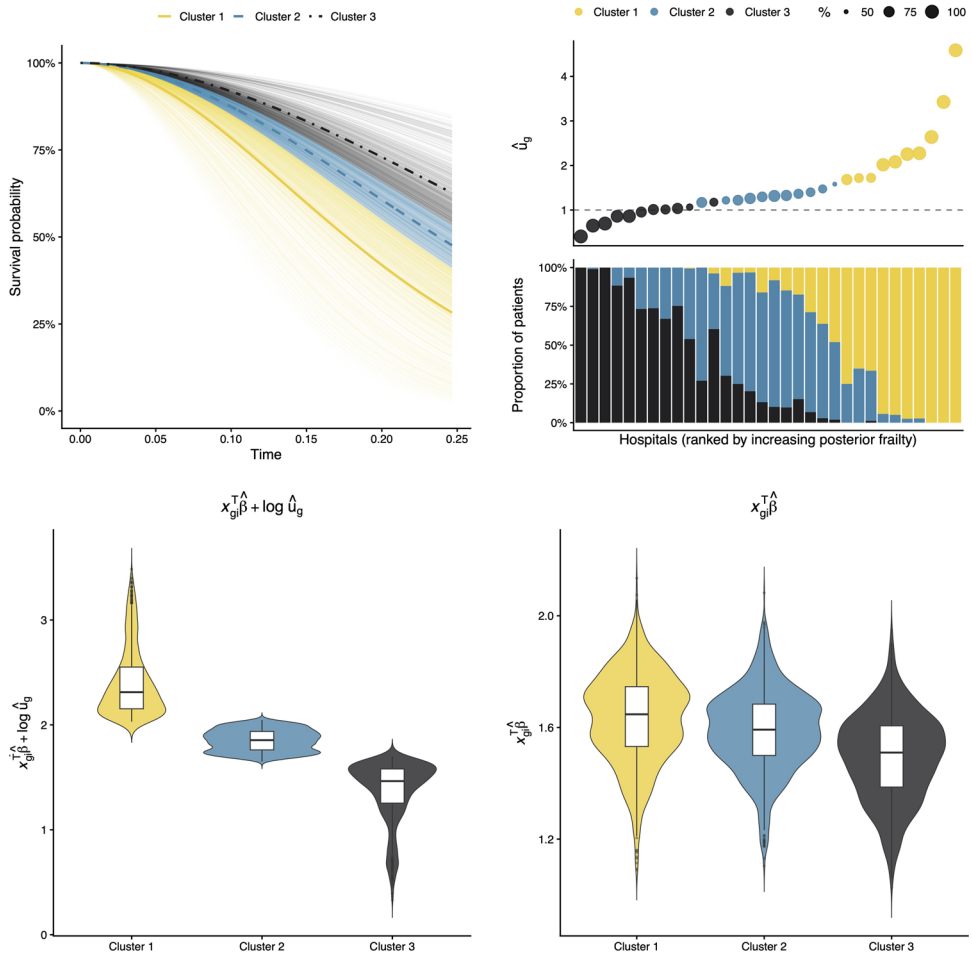


Figure 2 Case study results. a) Model-based estimated survival curves for each patient (thin lines), colored by cluster assignment, with the cluster specific average survival overlaid. b, top) Posterior frailty estimates $\hat{u}_g$ for each hospital, ranked by increasing frailty, colored by dominant cluster; point size reflects the percentage of patients belonging to the dominant cluster. b, bottom) Cluster composition of each hospital, ordered as in the top panel. c), d) Violin plots with overlaid boxplots showing the distribution of c) the total log-hazard contribution and d) the fixed-effects linear predictor only, across the 3 clusters.

The estimated survival curves in Fig. 2a reveal clear separation across the 3 clusters: Cluster 3 exhibits the most favorable and homogeneous prognosis, Cluster 1 the steepest decline and widest spread, and Cluster 2 occupies an intermediate position. Kaplan–Meier curves confirm this separation with a highly significant log-rank test ($P < 0.001$, Fig. S7.1). At the hospital level, Fig. 2b reveals a strong association between posterior frailty estimates $\hat{u}_g$ and dominant cluster membership. Hospitals predominantly composed of Cluster 3 patients concentrate on the left of the frailty ranking ($\hat{u}_g < 1$), indicating a protective institutional effect, whereas hospitals dominated by Cluster 1 patients appear on the right ($\hat{u}_g > 1$), reflecting excess risk beyond that explained by observed covariates. While intrahospital heterogeneity in cluster composition exists, a dominant cluster is consistently identifiable within each hospital: 13 hospitals (40.6%) showed strong single-cluster predominance ($\geq 80\%$ of patients), of which 5 were entirely associated with 1 cluster, while only 3 hospitals exhibited a truly mixed composition with no cluster exceeding 60%.

Finally, Fig. 2c and d illustrate the 2 sources of risk driving the stratification. The covariate-driven term $\boldsymbol{x}'_{gi}\hat{\boldsymbol{\beta}}$ already shows modest separation across clusters, with Cluster 1 patients carrying slightly

higher baseline risk; however, the distributions overlap considerably. Once the hospital-level frailty shift $\log \hat{u}_g$ is incorporated, the separation becomes sharp and nearly complete, confirming that hospital-level heterogeneity—captured by the frailty parameter θ—substantially amplifies the risk separation beyond observed covariates.

5. DISCUSSION

Motivated by the need to identify latent patient subgroups in hierarchical time-to-event data, in this work we propose a framework that jointly performs patient risk-based clustering and survival modeling while accounting for group-level dependence through a parametric shared frailty model, ensuring that cluster assignments are informed by underlying survival dynamics. From a methodological perspective, our approach addresses a gap in the existing literature by jointly modeling survival outcomes, latent patient subgroups, and group-level dependence within a single likelihood-based framework, allowing patient stratification to emerge from estimated risk profiles while properly accounting for within-hospital correlation.

The simulation study demonstrates that the proposed methodology reliably recovers latent cluster structures across a broad range of settings. In particular, the results indicate that cluster recovery is driven primarily by the strength of the underlying clustering signal—namely, the degree of separation between cluster-specific risk profiles and the balance of latent cluster sizes—whereas the magnitude of the frailty variance has only a limited impact on performance, as it is well captured by the model. Because the proposed clustering procedure relies on a parametric shared frailty model, the resulting similarity structure depends on the assumed frailty and baseline hazard distributions. Nevertheless, the simulation experiments show that the methodology remains robust under misspecification of both the frailty distribution and the baseline hazard function, with only limited deterioration in clustering performance even under joint misspecification, and with accurate estimation of the parameters of the parametric frailty model when penalization is appropriately tuned.

When applied to the Enhance-Heart cohort of HF patients hospitalized for COVID-19, the method identifies 3 clinically interpretable clusters with distinct risk profiles, differing significantly in age, comorbidity burden, respiratory conditions, and survival trajectories ($P < 0.05$), uncovering heterogeneity in event-time risk that would remain hidden under more standard approaches. Beyond patient-level stratification, the nonnegligible frailty dispersion parameter $\theta = 0.573$ suggests that latent hospital factors—such as resource availability, clinical protocols, or pandemic-specific pressures—influence survival outcomes beyond observed covariates. These findings support the joint use of clustering and frailty as complementary tools for disentangling patient- and hospital-level sources of heterogeneity in time-to-event data.

Some limitations warrant consideration. First, the construction of the similarity graph relies on the choice of the neighborhood size k, which influences graph sparsity and cluster formation. Although adaptive tuning of the penalization parameter λ embedded in the algorithm mitigates this issue in practice, the relationship between k and the target number of clusters C remains an area for further investigation. Furthermore, each unit is forced to connect to its k nearest neighbors regardless of whether these lie near the cluster core or at its boundary. This limitation has been noted in the spectral clustering literature (Cai et al. 2022), where pruning weak connections and modifying the optimization term through power-function asymptotics have been proposed as potential remedies. However, such approaches introduce alternative tuning parameters, such as the maximum neighborhood size, without yielding performance improvements in our setting.

Moreover, the proposed method does not explicitly account for confounding or informative selection mechanisms through causal inference techniques. The identified clusters should therefore be interpreted as *descriptive* risk-based stratifications conditional on the observed data and the assumed model, rather than as causally defined subgroups. Residual unmeasured heterogeneity is nevertheless partially mitigated through the shared frailty term, which absorbs latent variation linked to the hierarchical structure of the data, and through the restriction of the analysis to a relatively homogeneous population, which reduces heterogeneity in the selection mechanism.

In addition, in the current framework, a common baseline hazard is assumed across clusters, which is simplified in the penalization distance, while heterogeneity is captured through covariates and frailty terms. Future work will aim to extend the model to allow cluster-specific baseline hazards, enabling a more nuanced characterization of risk dynamics across patient subgroups, although this would imply the definition of a time-dependent similarity measure. This extension would require reconsidering the frailty structure—as different clusters would have distinct frailties—and is beyond the scope of the present work. Additional developments could explore alternative similarity measures (eg via survival functions), as well as extensions to semi-parametric or fully nonparametric baseline hazards, further increasing the model's flexibility.

Acknowledgments

This work is part of the *Enhance-Heart* project. Data were provided by "Unità Organizzativa Osservatorio Epidemiologico Regionale" and ARIA S.p.A. The authors acknowledge the support of MUR, grant Dipartimento di Eccellenza 2023-2027. Simulations were run on the Department of Mathematics HPC Cluster at Politecnico di Milano, funded by the same grant. We thank the associate editor and two anonymous reviewers for their helpful feedback, which helped us improve our manuscript.

Supplementary material

Supplementary material is available at *Biostatistics Journal* online.

Funding

A.R. and L.C. are funded by the National Plan for NRRP Complementary Investments "ANTHEM - AdvaNced Technologies for Human-centrEd Medicine" (PNC0000003).

Conflicts of interest

None declared.

Data availability

Software in the form of R code is available at https://github.com/alessandragni/HazClust. Data were provided by the Lombardy Region under permission and cannot be made publicly available for privacy reasons.

References

Abrahantes JC et al. 2007. Comparison of different estimation procedures for proportional hazards model with random effects. Comput Stat Data Anal. 51:3913–3930. https://doi.org/10.1016/j.csda.2006.03.009

Adeghate EA, Eid N, Singh J. 2021. Mechanisms of covid-19-induced heart failure: a short review. Heart Fail Rev. 26:363–369. https://doi.org/10.1007/s10741-020-10037-x

Ahlqvist E et al. 2018. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 6:361–369. https://doi.org/10.1016/S2213-8587(18)30051-2

Alteri C et al. 2021. Genomic epidemiology of sars-cov-2 reveals multiple lineages and early spread of sars-cov-2 infections in Lombardy, Italy. Nat Commun. 12:434. https://doi.org/10.1038/s41467-020-20688-x

Bader F, Manla Y, Atallah B, Starling RC. 2021. Heart failure and covid-19. Heart Fail Rev. 26:1–10. https://doi.org/10.1007/s10741-020-10008-2

Balan TA, Putter H. 2020. A tutorial on frailty models. Stat Methods Med Res. 29:3424–3454. https://doi.org/10.1177/0962280220921889

Buginga G, de Souza e Silva E. 2024. Clustering survival data using a mixture of non-parametric experts [preprint], arXiv, arXiv:2405.15934.

Cai Y, Huang JZ, Yin J. 2022. A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering. Neurocomputing. 493:191–203. https://doi.org/10.1016/j.neucom.2022.04.030

Caldera L et al. 2025. Cluster-weighted modeling of lifetime hierarchical data for profiling covid-19 heart failure patients [preprint], arXiv, arXiv:2507.12230.

Cavinato L et al. 2023. Explainable domain transfer of distant supervised cancer subtyping model via imaging-based rules extraction. Artif Intell Med. 138:102522. https://doi.org/10.1016/j.artmed.2023.102522

Chapfuwa P, Li C, Mehta N, Carin L, Henao R. 2020. Survival cluster analysis. In: Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL '20). New York, NY, USA: Association for Computing Machinery, 60–68. https://doi.org/10.1145/3368555.3384465

Chen D et al. 2016. An algorithm for creating prognostic systems for cancer. J Med Syst. 40:160. https://doi.org/10.1007/s10916-016-0518-1

Collins FS, Varmus H. 2015. A new initiative on precision medicine. N Engl J Med. 372:793–795. https://doi.org/10.1056/NEJMp1500523

Corrao G et al. 2017. Developing and validating a novel multisource comorbidity score from administrative data: a large population-based cohort study from Italy. BMJ Open. 7:e019503. https://doi.org/10.1136/bmjopen-2017-019503

Duchateau L, Janssen P. 2008. The frailty model. Springer.

Fan K. 1949. On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc Natl Acad Sci USA. 35:652–655. https://doi.org/10.1073/pnas.35.11.652

Guo X. 2015. Robust subspace segmentation by simultaneously learning data representations and their affinity matrix. In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15). AAAI Press, 3547–3553.

Hubert L, Arabie P. 1985. Comparing partitions. J Classif. 2:193–218. https://doi.org/10.1007/BF01908075

Janke AT et al. 2021. Analysis of hospital resource availability and covid-19 mortality across the United States. J Hosp Med. 16:211–214. https://doi.org/10.12788/jhm.3539

Kaufman L, Rousseeuw PJ. 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Liu C et al. 2022. Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data. IEEE/ACM Trans Comput Biol Bioinform. 19:1193–1202. https://doi.org/10.1109/TCBB.2020.3010509

Manduchi L et al. 2021. A deep variational approach to clustering survival data. In International Conference on Learning Representations (ICLR 2022). https://openreview.net/forum?id=RQ428ZptQfU

Mohar B, Alavi Y, Chartrand G, Oellermann O. 1991. The Laplacian spectrum of graphs. Graph Theory Combinatorics Appl. 2:12.

Munda M, Rotolo F, Legrand C. 2012. parfm: Parametric frailty models in r. J Stat Soft. 51:1–20. https://doi.org/10.18637/jss.v051.i11

Nie F, Wang X, Huang H. 2014. Clustering and projected clustering with adaptive neighbors. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). New York, NY, USA: Association for Computing Machinery, 977–986. https://doi.org/10.1145/2623330.2623726

Nie F, Wang X, Jordan MI, Huang H. 2016. The constrained Laplacian rank algorithm for graph-based clustering. AAAI 2016. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. AAAI Press. p. 1969–1976.

Tosado J et al. 2020. Clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction. Sci Rep. 10:3811. https://doi.org/10.1038/s41598-020-60140-0