POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Riccardo Scaramuzza

Advisor: Prof. Francesca Ieva

Co-advisors: Dr. Chiara Masci, Dr. Marta Spreafico

Academic year: 2020-2021

1. Introduction

This thesis faces the problem of joint modelling of hospitalizations and survival of patients affected by Heart Failure, with a focus on the effect that a pharmacological treatment based on ACE Inhibitors has on these two processes.
Heart Failure is a chronic condition consisting in the deterioration of the function of a patient's heart. Its study is considered to be of primarily importance, due to its prevalence and impact over the sanitary system.
We investigate Survival analysis tools able both to model two correlated processes, the former regarding recurrent events (i.e. hospitalizations) and the latter terminal ones (i.e. deaths), and to assess the effect that exogenous variables (e.g. ACE inhibitors therapy) have on them. We identify frailty models [4], which are Cox models in which a random effect is added to the linear predictor, as a suitable tool both in a recurrent and terminal events framework. Moreover, their application allows the simultaneous modelling of the two processes through the linking of the hospitalizations and death frailties [5]. Our main contribution is represented by an innovative approach extending the state of the art joint model proposed by Ng et al. in [3], in which the two processes' frailties follow a bivariate non parametric discrete distribution. This frailty formulation reveals a big potential from an interpretative point of view, especially for the application at hand. In fact, it enables a more direct analysis of the induced partition of patients in subpopulations characterized by different levels of fragility, which can be easily translated in a providers' assessment. We finally provide a comparative study to verify the effectiveness of our model in a controlled setting.

2. ACE Inhibitors Dataset

We consider data coming from an administrative database of Regione Lombardia, which records clinical courses and pharmacological prescriptions of subjects affected by Heart Failure. We focus on patients who undergo an ACE inhibitors treatment in the period from January 1st, 2006 to December 31st, 2012. For each patient, the index date coincides with the discharge after the first hospitalization due to Heart Failure. We adopt a *gap times* timescale, i.e. each patients clinical history is declined in repeated observations, characterized by a time-to-event variable **GapEvent** which expresses the days elapsed from the previous patient's hospitaliza-

tion to the next one. The last gap time of each patient expresses the time elapsed from the last known hospitalization to the terminal event, which may be death or censoring. The nature of each event is kept track of through two dummy variables, respectively **Event** and **Death**.
To assess the effect of the considered ACE inhibitors treatment on survival and hospitalizations, we extend the approach proposed in [1], designing a time dependent binary classifier for adherent subjects (variable **Adherent**). At each event in a patient's history, we compute the proportion of days covered by prescriptions of ACE inhibitors since the patient index date; then, if this proportion exceeds a threshold of 80% the patient is considered adherent to treatment, otherwise not.
Moreover, each entry in the dataset comprehends two time-dependent variables, **AgeEvent** and **Comorbidity**, which respectively indicate the age and the number of known comorbidities of a patient at the beginning of the corresponding gap time. Finally, the last variable included in the modelling is the patient gender (**Sex**).
Table 1 reports as an example the data table of a patient in the ACE inhibitors dataset.

3. Methods

In our context, we need a tool to model the effect of exogenous variables on possibly censored time-to-event outcomes regarding the hospitalizations and death processes, taking into account the heterogeneous fragilities of patients.

3.1. Frailty Models

We initially model the two processes separately through Cox proportional hazard *frailty models* [4]. They express the hazard (i.e. the probability of experiencing an event at time t) similarly to the Cox model, but they exploit the introduction of a random unobserved covariate (the frailty) that describes the heterogeneity at patient level unexplained by the observed set of covariates. In our case, they can be applied to account for within-subject correlated times, which are now assumed to be independent conditionally on the covariate vector and on the unobserved random effects. In particular, we express the hospitalization and death hazards for each patient $i, i = 1, ...N$, as follows

$$\begin{aligned} h_i^R(t|\boldsymbol{x}_i^R(t)) &= h_0^R(t)\exp\{\boldsymbol{\beta^T}\boldsymbol{x}_i^R(t) + u_i\} \\ h_i^D(t|\boldsymbol{x}_i^D(t)) &= h_0^D(t)\exp\{\boldsymbol{\gamma^T}\boldsymbol{x}_i^D(t) + v_i\} \end{aligned} \quad (1)$$

where t refers to a gap time with respect to the last known hospitalization event; h_0^R and h_0^R are the hospitalization and death baseline hazard functions, respectively; $\boldsymbol{x}_i^R(t)$ and $\boldsymbol{x}_i^D(t)$ are the observed covariates at time t; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the estimated coefficients of the two models; u_i and v_i are patient-specific additive frailties, which follow two independent Normal distributions, centered in zero and characterized by their variances parameters

$$\begin{aligned} p(u) &= N(0, \theta_u^2) \\ p(v) &= N(0, \theta_v^2). \end{aligned} \quad (2)$$

We fit the model to our data using the R package `coxme`, which implements the estimation procedure proposed in [4].

3.2. Joint Models

Joint frailty models allow to study the joint evolution over time of our two correlated survival processes by linking the two processes' frailties. In our work we consider at first the model proposed by Rondeau et al. [5] in 2007. It comprehends a single random frailty, η, which is normally distributed, but acts differently on the two processes' hazards through the parameter α. Following the notation adopted in Equation 1, the hazards are modelled as follows

$$\begin{cases} h_i(t|\eta_i, \boldsymbol{x}_i^R(t)) = h_0^R(t)\exp\{\eta_i + \boldsymbol{\beta}^T\boldsymbol{x}_i^R(t)\} \\ h_i(t|\eta_i, \alpha, \boldsymbol{x}_i^D(t)) = h_0^D(t)\exp\{\alpha\eta_i + \boldsymbol{\gamma}^T\boldsymbol{x}_i^D(t)\} \end{cases} \quad (3)$$

The model is fitted to data through the R package `frailtypack`.
Then, we consider a model proposed by Ng et al.[3] in 2020, where the hazards are modeled as in Equation 1, but the two processes' frailties are jointly modelled as a bivariate Normal distribution

$$p\big([u_i, v_i]|\boldsymbol{\mathcal{E}}\big) = \mathcal{N}_2(\mathbf{0}, \boldsymbol{\mathcal{E}}) \quad (4)$$

where $\boldsymbol{\mathcal{E}}$ stands for

$$\boldsymbol{\mathcal{E}} = \begin{bmatrix} \theta_u^2 & \rho\theta_u\theta_v \\ \rho\theta_u\theta_v & \theta_v^2 \end{bmatrix} \quad (5)$$

ID	Sex	Adherent	AgeEvent	Comorbidity	GapEvent	Event	Death
10003004	F	0	75	5	229	1	
10003004	F	1	75	6	131	1	
10003004	F	0	76	6	168	1	
10003004	F	0	77	7	353	1	
10003004	F	1	79	7	1,153	0	1

Table 1: Data table of patient 10003004.

This formulation allows a well-defined interpretation of all the parameters involved, being θ_u^2 and θ_v^2 the quantifiers of unobserved heterogeneity in the two processes, while ρ models their dependence. In [3], the authors propose aswell an innovative estimation routine for the model, which has not been implemented in a dedicated R package yet. Thus, we develop our custom implementation to fit the model to data.

3.3. Joint Discrete Nonparametric Frailty Model

Our main contribution consists in the development of a model which assumes the same shape for the hospitalization and death hazards as in Equation 1, but a bivariate non parametric discrete distribution for the frailties. This choice stems from the fact that a discrete distribution of frailties is easier to understand and may translate in a providers' assessment. We design and implement a specific EM algorithm for the model's training, which was inspired by [2].
According to our formulation, random effects u_i and v_i are distributed according to P^*, which is an unknown measure on $\mathbb{R}^2$

$$[u, v]_i \overset{iid}{\sim} P^* \quad \forall i = 1..N \tag{6}$$

Such measure in supposed to be discrete and with a finite support, thus it can be characterized by a vector of points in $\mathbb{R}^2$, $\boldsymbol{P}$, and a vector of weights, $\boldsymbol{w}$. Notice that each weight expresses the probability of a patient to be assigned to a certain point l, $l = 1, ..., L$ and thus the sum of the weights is constrained to be unitary. Moreover, the number of points constituting the support of the distribution, L, is assumed to be unknown a priori.
Initially considering this parameter as fixed, we can write each patient's contribution to the likelihood of the model as a mixture of L components. Each component coincides with the product of the individual contributions (relative to patient i) to the full loglikelihood of two independent Cox models with fixed intercept, modelling respectively the recurrent and terminal event process

$$L(\boldsymbol{\Omega}; data|z_{il}) = \prod_{l=1}^{L}\prod_{i=1}^{N}\left[L_i^{full}(\boldsymbol{\Omega}; data|[u, v]_i = P_l)\right]^{z_{il}}. \tag{7}$$

The abscissa and ordinata of each point $\boldsymbol{P}_l$ specify, respectively, the fixed intercept of the recurrent and terminal event Cox models, while z_{il} are a set of binary auxiliary random variables indicating if a patient i is assigned to the point l. In order to obtain estimates for $\boldsymbol{\Omega} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{P}, h_0^R(t), h_0^D(t)]$, we design a specific EM algorithm, in which at each iteration the model likelihood is firstly averaged with respect to the z_{il} variables and then maximized.
The EM algorithm is then generalized to cope with an a priori unknown number of support points L through its integration into a wrapper support reduction procedure. The first step consists in the definition of a grid of points in $\mathbb{R}^2$, which ideally covers the region in which the support of the discrete distribution is believed to lie. We evaluate two methods: the former involves sampling an high number of points from a bivariate Normal distribution, whose parameters are set according to previous knowledge (e.g. looking at estimates of the disjoint or Ng et al. model), and initializing their weights according to the corresponding Normal density; the latter consists in defining a uniform distribution over a rectangle in $\mathbb{R}^2$, whose boundaries are set still according to available knowledge. We decide to adopt the second method, as it is less informative and results more robust with respect to randomization. At the start of each iteration, we merge the minimum distance couple of points in the actual grid whose Euclidean distance is less

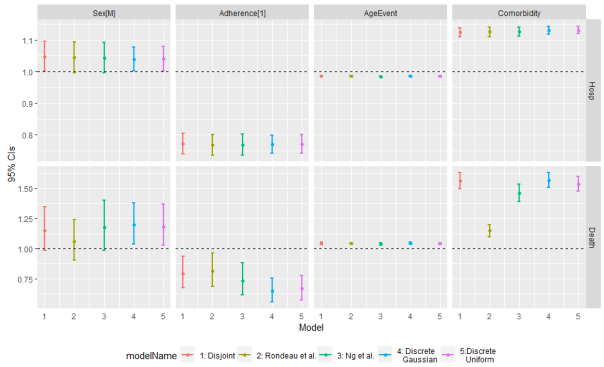

Figure 1: Comparison of estimated hazard ratios and their 95% CI in the trained models. Considered models are: disjoint (pink), Rondeau et al. (pistachio green), Ng et al. (teal), Discrete Nonparametric Frailty with Gaussian Initialization (light blue) and Discrete Nonparametric Frailty with Uniform Initialization (purple).

Variables	Estimate	StdDev	HR	CI95	pvalue
Recurrent Events					
Sex [M]	0.039	0.019	1.039	[1.003,1.079]	0.034
Adherent [1]	-0.259	0.019	0.771	[0.743,0.800]	<2e-16
AgeEvent	-0.015	0.001	0.985	[0.984,0.987]	<2e-16
Comorbidity	0.123	0.005	1.131	[1.119,1.143]	<2e-16
Recurrent Events					
Sex [M]	0.169	0.073	1.184	[1.025,1.366]	0.021
Adherent [1]	-0.407	0.078	0.665	[0.571,0.755]	1.7e-07
AgeEvent	0.039	0.004	1.041	[1.032,1.049]	<2e-16
Comorbidity	0.429	0.020	1.535	[1.476,1.597]	<2e-16
Frailty	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**
u	-0.466	-0.194	0.079	0.231	0.468	0.679
v	-1.872	-0.859	-0.090	1.166	2.277	3.088
w	0.208	0.234	0.206	0.217	0.071	0.063

Table 2: Summary of the Nonparametric Discrete Frailty model with Uniform initialization. For categorical variables, the considered stratum is indicated between brackets. Points of the identified frailty discrete distribution are characterized through their abscissa (u), ordinata (v) and weight (w).

than a threshold (*MinDist*). The new point is simply defined as the median of the old ones connecting segment, while its weight as the sum of the old points ones. The procedure is repeated until any couple of points distance is under the threshold. The last step consists in the deletion of eventual masses to which no patients are assigned in the latent partition extracted after the maximization step. The algorithm stops when a given number of iterations is reached or the number of masses in the discrete distribution is stable (no reduction happens in the current iteration) and the difference between old and updated weights of the discrete distribution, computed in maximum norm, is less than 1e-03.

4. Results

In the fitting of all models we consider for both processes the whole set of available covariates (**Sex**, **Adherence**, **AgeEvent** and **Comorbidity**). The nonparametric discrete frailty models are fitted using a *MinDist* value of 0.25, which is initially believed to be suitable to spot significantly different fragility classes of patients. The trained models are compared from two perspectives: coefficients' estimation and random effects' characterization.

From the coefficients' estimation point of view, we expect all different models to yield consistently similar values. In Figure 1 are visualized the Hazard Ratio estimates and their respective 95% confidence intervals, which confirm this expected result.

Actually, the only significant difference regards the comorbidity coefficient estimated by Rondeau et al. model, which is likely to be an error due to numerical instability in its estimation routine. As reference, we analyze in the following the estimates provided by the discrete nonparametric frailty model with Uniform initialization (see Table 2).

Looking at the hazard ratio of the covariate **Sex**, male subjects are suggested to be slightly more prone to risk of hospitalization (HR=1.035) and death (HR=1.197).

The covariate **Adherent** results statistically significant at any level for the two processes in all trained models. According to the Uniform initialization discrete nonparametric model, being adherent yields a 22.9% decrease in the risk of a new hospitalization (HR=0.771) and a 33.5% decrease in the risk of death. From a clinical point of view, such results finally endorse the efficacy of the ACE inhibitors treatment for heart failure, as it leads to a significant reduction of the hospitalizations rate (and thus of critical HF events) of adherent patients during their clinical path, in addition to increasing their survival probability.

The covariate **AgeEvent** results, for both processes, statistically significant at all levels in all trained models. Its effect on the hospitalization hazard is a 1.5% reduction of the risk of hospitalization per year (HR=0.985), while on the death hazard it yields an increase of the risk of 4.1% (HR=1.040). Clinically speaking this can be explained by the fact that part of the risk of experiencing a new hospitalization

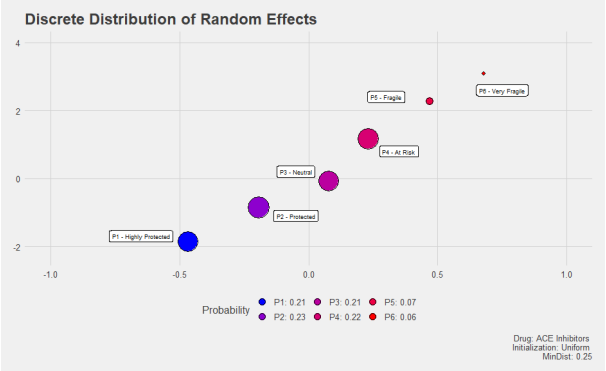

Figure 2: Discrete Distribution of Random Effects. The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is representative of the probability of a patient to belong to the corresponding latent population.

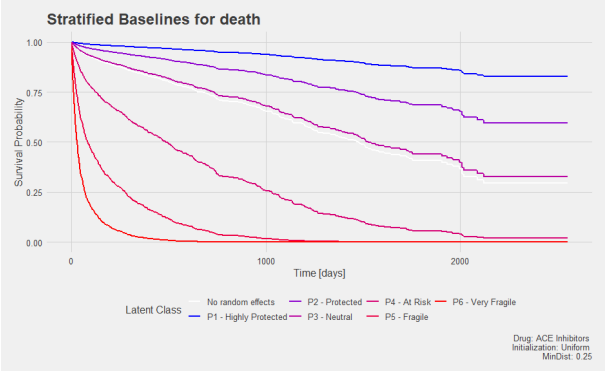

Figure 3: Stratified Survival Probability Baseline curves of the terminal event process, associated to the discrete distribution of random effects identified in the Uniform Initialization model. The color of each curve is the same of the corresponding random effect point as in Figure 2. The white line represents the survival probability baseline curve of the model without random effects.

is replaced by the risk to die when patients get older. This is reasonable especially in our case, as we are considering elderly persons (median age at first hospitalization of 74, IQR=[66;80]), but is likely to be different when young subjects are involved.
The covariate **Comorbidity** results to be in all models statistically significant for both the processes. It yields an increase of 13.1% in the risk of hospitalization and a very high increase of 53.5% in the risk of death per comorbidity registered. The role comorbidities have in increasing the mortality and hospitalizations of heart failure patients is well documented in medical literature and confirmed by our analysis.

From the frailties characterization point of view, the disjoint model yields estimates of $\theta_u^2 = 0.093$ and $\theta_v^2 = 0.428$ for, respectively, the hospitalization and death random effects' variances. Generally, the fact that the within-patient variability is very low in the hospitalizations process, while is quite high in the death one, may be due to the fact that the two processes show different timescales and are trained on very different amounts of data. (15,978 against 3,232). This difference may also be explained from a clinical point of view, as subjectivity is likely to be more relevant on mortality that on hospitalizations, which are regulated by fixed procedures.
As mentioned in [3], when the frailties' are not jointly modelled the heterogeneity involved in two correlated processes is likely to be underestimated. In Rondeau et al. model, the estimated variance of the random effect η is slightly higher than its disjoint model counterpart (0.114 against 0.093). The estimated α parameter, instead, is 2.660: since it acts multiplicatively on the random effect (see Equation 3), it yields a variance of 0.799, which is significantly higher than the independent model one (0.428). However, this joint frailty formulation seems too simplistic, in addition to the fact that the α parameter has not a clear interpretation.
Ng et al. model yields estimates of $\theta_u^2 = 0.124$ and $\theta_v^2 = 1.378$ for the hospitalization and death frailties' variances. This magnification effect (with respect to the estimates provided by the disjoint model) is likely to be due to the very strong dependence between the two processes that the model identifies: the estimate for the correlation parameter is $\rho = 0.879$, which suggests that in our cohort patients which are naturally more prone to the risk of a new hospitalization are also naturally more prone to the risk of death. However, even if highly correlated, the strong difference between the two processes random effects variances suggests a far more important role of randomicity in the terminal event process.
Our nonparametric frailty model identifies the discrete distribution reported in the third section of Table 2 and visualized in Figure 2. The estimated discrete distribution consists in six points disposed in a diagonal pattern, whose range is consistent with the variance estimates obtained in Ng et al. model, and show left skewness: point **P1** and **P2**, asso-

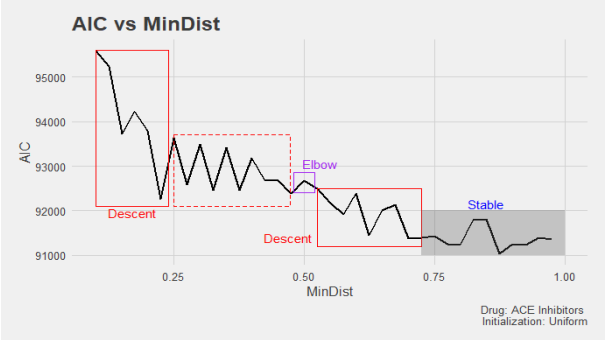

Figure 4: AIC curve as function of the MinDist parameter, computed through the Uniform initialization discrete nonparametric frailty model.

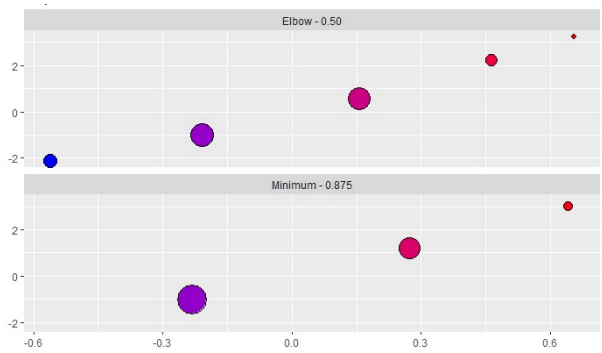

Figure 5: Discrete distribution of random effects related to relevant values of MinDist spotted analyzing Figure 4

ciated with a probability of 21% and 23%, identify respectively a *Highly Protected* and a *Protected* subpopulation; Point **P3** is related to a subpopulation *Neutral* to random effects, with almost a 20% probability for a patient to belong to it; Point **P4** identifies a relevant group of patients (*At Risk*) slightly more prone to the risk of a new hospitalization and death, with a probability of 20.5%. Point **P5** and **P6** identify two outlier subpopulations of *Fragile* and *Very Fragile* patients, associated with small probabilities (respectively, 7% and 6%). To quantify the influence of the identified frailties discrete distribution on the two processes we look at the induced stratified baseline survival curves. As an example, Figure 3 reports the terminal event process induced stratified baselines. We note that they show very different profiles in terms of survival, in particular the *Fragile* and *Very Fragile* subpopulations (red curves), whose subjects are likely to depart within a short time.

Finally, we investigate the change in the *MinDist* threshold, which turns out to be the main factor influencing the final distribution discovered. We look at the dependence between the parameter to be tuned and a simple fitting criterion, the Akaike Information Criterion, in order to identify promising candidates. In Figure 4 is reported the curve obtained considering a set of 37 values for *MinDist* ranging from 0.1 to 1, which are apart from each other of 0.025. We choose $MinDist = 0.875$, where the curve achieves its minimum, as the most promising candidate. It yields the discrete distribution reported in bottom panel of Figure 5, which comprehends a main *Protected* subpopulation (purple point, probability of 54.9%) and another relevant *At Risk* subpopulation (magenta point, probability of 36%), in addition to an outlier point representative of *Very Fragile* patients (small red point, probability of 9%).

5. Conclusions

Our innovative approach results to be an effective inferential tool to jointly model hospitalizations and survival of patients. It yields covariate coefficients estimates consistent with models used in literature, providing an easy to understand but richer frailty characterization, as it enables further analyses. For example, it allows us to argue that the assessed positive effect that adherence to the ACE inhibitors treatment have in reducing the hospitalization and death rates is likely to be unappreciable for outlier patients belonging to the identified *Very Fragile* class, due to their personal fragility.

References

[1] Andrade, S. et al. Methods for evaluation of medication adherence and persistence using automated databases. *Pharmacoepidemiology and Drug Safety*, 15:565–574, 2006.

[2] Gasperoni, F. et al. Non-parametric frailty cox models for hierarchical time-to-event data. *Biostatistics*, 2020.

[3] Ng, S.K. et al.a. Joint frailty modelling of time-to-event data to elicit the evolution pathway of events: A generalised linear mixed model approach. *Biostatistics*, 0(0):1–25, 2020.

[4] Ripatti, S. et al. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56:1016–1022, 2002.

[5] Rondeau, V. et al. Jointfrailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *International Journal of Epidemiology*, 8:708–721, 2007.