MOX-Report No. 90/2023

Dynamic treatment effect phenotyping through functional survival analysis

Gregorio, C.; Baj, G.; Barbati, G.; Ieva, F.

DYNAMIC TREATMENT EFFECT PHENOTYPING THROUGH FUNCTIONAL SURVIVAL ANALYSIS

Caterina Gregorio[1,2], Giovanni Baj [3], Giulia Barbati [2], Francesca Ieva [1,4]

[1] MOX - Modelling and Scientific Computing, Department of Mathematics Politecnico di Milan,
Piazza Leonardo Da Vinci 32, Milan 20133, Italy,

[2] Biostatistics Unit, Department of Medical Sciences, University of Trieste, Via Valerio 4\1, Trieste 34100, Italy,
Piazza Leonardo Da Vinci 32, Milan 20123, Italy,

[3] Department of Mathematics and Geosciences, University of Trieste, Via Valerio 6, 34100, Trieste, Italy,

[4] HDS, Health Data Science center, Human Technopole, Viale Rita Levi Montalcini 1, Milan 20157, Italy.

ABSTRACT

In recent years, research interest in personalised treatments has been growing. However, treatment effect heterogeneity and possibly time-varying treatment effects are still often overlooked in clinical studies. Statistical tools are needed for the identification of treatment response patterns, taking into account that treatment response is not constant over time. We aim to provide an innovative method to obtain dynamic treatment effect phenotypes on a time-to-event outcome, conditioned on a set of relevant effect modifiers. The proposed method does not require the assumption of proportional hazards for the treatment effect, which is rarely realistic. We propose a spline-based survival neural network, inspired by the Royston-Parmar survival model, to estimate time-varying conditional treatment effects. We then exploit the functional nature of the resulting estimates to apply a functional clustering of the treatment effect curves in order to identify different patterns of treatment effects. The application that motivated this work is the discontinuation of treatment with Mineralocorticoid receptor Antagonists (MRAs) in patients with heart failure, where there is no clear evidence as to which patients it is the safest choice to discontinue treatment and, conversely, when it leads to a higher risk of adverse events. The data come from an electronic health record database. A simulation study was performed to assess the performance of the spline-based neural network and the stability of the treatment response phenotyping procedure. In light of the results, the suggested approach has the potential to support personalized medical choices by assessing unique treatment responses in various medical contexts over a period of time.

1 Introduction

The field of personalized treatments has been experiencing a noticeable surge in research interest. This can be attributed to the growing recognition that the one-size-fits-all approach has significant limitations. Nonetheless, despite this heightened focus, there remains a tendency to neglect the potential variations in treatment effects over time when considering time-to-event outcomes. When studying the problem of obtaining personalized treatment decisions in order to minimize the subject's risk of adverse events, it is important to acknowledge how time affects the response to the treatment. In fact, the treatment response that we observe is typically the result of multiple biological mechanisms that a treatment triggers in an individual. These may differ in their timing, magnitude and nature, as some may be beneficial and others may be side effects. Whereas there is a growing literature on optimal treatment rules using machine-learning algorithms such as reinforcement learning [1], the aim of this work is to provide a method to identify common treatment effect patterns over time to characterize the dynamic response of individuals to treatments. The rationale is that such a method can 1) provide valuable information on the whole time horizon of interest without restricting the attention to a single point in time and, 2) inform on the timing and direction of the treatment response of individuals. As a result, the medical expert can make a judgment of the best treatment decision to take for a specific patient profile. For example, this approach can help distinguish subject profiles that have an early vs. late response to a treatment,
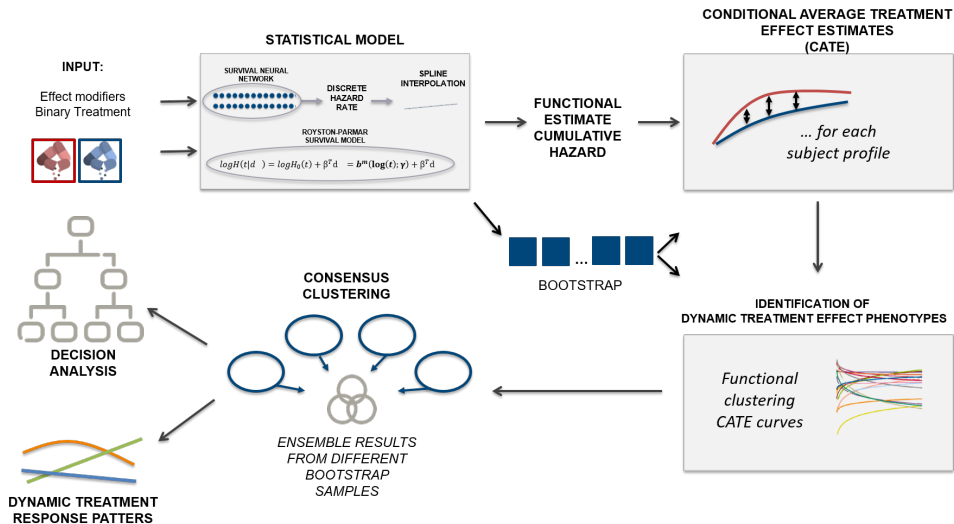
Figure 1: Summary of the proposed method to obtain dynamic treatment effect phenotypes.

or it can provide evidence on whether the effect of a treatment or exposure tends to fade away or increase over time. According to the specific medical context and specific decision problem, this information can have different implications, and it is important for the decision maker to have them before choosing one treatment alternative over the other. In order to do so, the Average Treatment Effect conditioned on the effect modifiers, the so-called Conditional Average Treatment Effect (CATE), is estimated taking into account its possibly time-varying nature. To characterize the different treatment-effect phenotypes over time, a clustering procedure is performed considering the time dimension of the CATE. Functional Data Analysis (FDA)[2] is a part of statistics that collects methods to model data that can be thought of as being realizations of functional random variables. Usually, starting from discrete observations, the first step consists of obtaining functional estimates for each unit using smoothing. In survival analysis, the Royston-Parmar survival model[3] allows retrieving proper functional estimates of the survival curves by modelling the baseline hazard with natural cubic splines. The spline-based model allows obtaining smooth estimates for the cumulative hazard and survival functions, meaning that derivatives, i.e. hazard, are also available. Similarly to some FDA methods, smoothing is achieved with basis expansion, which allows controlling for the degree of smoothing without imposing strict functional forms. Moreover, the model can also be extended to include time-dependent effects. The main limitation of this method is that model selection is challenging when possible interactions among covariates and time-dependent effects are present. As an alternative, we propose to use a survival neural network approach to estimate the CATEs enabling us to easily capture all relevant interactions and time-dependent effects. Among the different methods, we consider the Logistic-Hazard method, also called Nnet-survival[4]. It is a discrete-time method that models the hazard function non-parametrically. Inspired by the Royston-Palmar model and FDA, we propose to interpolate the (discrete) neural network's predictions with natural cubic splines, to obtain smooth estimates of CATEs.

In FDA, a natural way to group and classify curves in an unsupervised manner is functional clustering. Here, we used it to identify relevant phenotypes that describe different responses to treatment over time, which can then be mapped to the different subjects' profiles.

The application that motivated this work is the discontinuation of treatment with Mineralocorticoid receptor Antagonists (MRAs) in patients with heart failure, where there is no clear evidence as to which patients it is the safest choice to discontinue treatment and, conversely, when it leads to a higher risk of adverse events. The data come from an Italian Electronic Health Record Database. The aim of the study is to provide evidence for the development of individualised recommendations for the discontinuation of MRAs according to the relevant clinical characteristics of the subjects. In Section 2 we describe the statistical methods. Section 2.1 introduces the notation and the theoretical framework. Details of how CATE can be estimated through either the spline-based model or the spline-based survival neural network are given in Section 2.2. The clustering procedure and the identification of the different phenotypes are reported in Section 2.3. Section 3 is devoted to the simulation study, whereas the description of the data used, the motivating problem and the results of the application are reported in Section 4. Finally, discussion and conclusions are reported in Section 5. A summary of the different steps of the method is reported in Figure 1.

2 Methods

2.1 Notation and Formal Framework

We let D and C the event of interest and censoring time respectively. We assume that C is non-informative with respect to the event time. For each subject, we only observe the couple $min(D, C)$, $\delta = \mathbb{1}(D \leq C)$. The binary random variable indicating the covariate of interest, i.e., the treatment or exposure, is denoted by Z. Let $\mathbf{X}$ be a p-dimensional vector of observable covariates with $p > 1$.

We define as $D^{(1)}$ the potential outcome if the treatment/exposure is received and $D^{(0)}$ the corresponding potential outcome without treatment/exposure.

We are interested in estimating the Conditional Average Treatment Effect (CATE) at each time $t \in [0, w]$ on a subset of the covariate vector $\mathbf{X}$, denoted as $\mathbf{X}_1$, that contains the treatment-effect modifiers :

$$\tau_{\mathbf{x}_1}(t)^{[0,w]} = E\{f[g(D^{(1)}; t), g(D^{(0)}; t)]|\mathbf{X}_1 = \mathbf{x}_1\} \tag{1}$$

where $g(\cdot)$ can be either the survival, hazard or cumulative hazard function and $f(\cdot)$ is a measure of effect, typically either the ratio or the difference. In the following, simplifying the notation, we omit the reference to the specific time horizon of interest, denoted as $[0, w]$, in $\tau_{\mathbf{x}_1^{[0,w]}}(t)$.

To identify the CATE, we need to assume the treatment groups are *conditionally exchangeable*:

$$[D^{(1)}, D^{(0)}] \perp Z|\mathbf{X}_1 \tag{2}$$

This assumption is also known as "ignorable treatment assignment". If the observed covariates contained in $\mathbf{X}$ but not in $\mathbf{X}_1$ act as confounders, then:

$$[D^{(1)}, D^{(0)}] \perp Z|\mathbf{X} \tag{3}$$

In the latter case, to ensure conditional exchangeability, a propensity score method such as matching or Inverse Probability of Treatment Weighting needs to be applied.

The steps of the method to identify treatment-effect phenotypes consist of 1) estimating the CATE as in equation 1, 2) identifying the grouping structure of the CATE curves through an unsupervised clustering method 3) characterizing the different behaviours in terms of response to the treatment 4) mapping the different subject profiles to their corresponding treatment-response behaviour.

2.2 Estimation of functional CATEs

In the Royston-Parmar Survival Model (R-PSM), the logarithm of the baseline cumulative hazard function is modelled as a natural cubic spline function of log time:

$$logH(t|\mathbf{d}) = logH_0(t) + \boldsymbol{\beta}^T\mathbf{d} = \mathbf{b}^m(log(t); \boldsymbol{\gamma}) + \boldsymbol{\beta}^T\mathbf{d} \tag{4}$$

where $\mathbf{b}^m(t; \boldsymbol{\gamma})$ is a natural cubic spline with m knots and $\mathbf{d}^T = [z \ \mathbf{x}_1 \ z \times \mathbf{x}_1]^T$. The flexibility in the form of the baseline hazard is given by the number of internal knots. The inclusion of interactions between covariates and the treatment variables enables the presence of heterogeneity in the treatment effects. This model can also be easily extended to include time-varying coefficients by modelling the spline coefficients in the function of the covariates for which we want a time-varying effect. The general model with time-dependent effect can be written as:

$$logH(t|d) = \gamma_0 + \boldsymbol{\gamma}\mathbf{b}^m(log(t)) + \boldsymbol{\beta}^T\mathbf{d} \tag{5}$$

where $dim(\boldsymbol{\gamma}) = m + 1$ and the jth component of $\boldsymbol{\gamma}$ is:

$$\gamma_j = \begin{cases} \gamma_{j0} & k = 0 \text{ Proportional Hazard model} \\ \gamma_{j0} + \sum_{r=1}^{k} \gamma_{jr}d_r & k \geq 1 \text{ Non-Proportional Hazard model for } d_1, \cdots d_r \end{cases} \tag{6}$$

The parameters in the model can be estimated using Maximum Likelihood (ML) and their uncertainty can be evaluated using standard ML asymptotic theory. The degree of freedom of the splines is typically chosen with the AIC or BIC and through visual comparison of the fitted hazard vs. non-parametric estimates. This model can be fitted in R using the package flexsurv [5] or stpm2 [6].

A model-free approach consists of using a survival neural network that estimates the hazard function. Specifically, we consider the Nnet-Survival method, also called Logistic-Hazard, which parametrizes the discrete-time hazard rate with a neural network and optimizes a survival likelihood expressed in terms of the discrete hazards non-parametrically [4]. In the Nnet-Survival method, follow-up time is divided into h intervals which are left-closed and right-open. The contribution of a generic time interval j to the overall log-likelihood is:

$$\sum_{i=1}^{d_j} = log(h_j^i) + \sum_{i=d_j+1}^{r_j} log(1 - h_j^i) \tag{7}$$

where r_j is the number of subjects at risk before the beginning of the interval, d_j is the number of subjects experiencing the event during the interval and, h_j^i is the hazard for an individual i during the time-interval j. The loss function in Equation 7 comes from classic discrete-time survival models, and its use in a neural network context is well justified by survival analysis theory. Furthermore, it naturally incorporates a time-varying baseline hazard rate and time-varying effect, since each time interval output node is fully connected to the last hidden layer's neurons. As architecture, we used a fully connected network with 2 hidden layers of 32 units each, and ReLU as a non-linear activation function. The neural network gives, for each subject profile, a h-dimensional output corresponding to a discrete set of hazard rates, one for each interval. We then propose to interpolate the hazard curves using natural cubic splines. In this way, similar to the previous model we obtain a smooth estimate of $H(t|d)$ that depends in this case on a non-linear function of the covariates x_1 and the treatment indicator z. This neural network was implemented in python using pycox [7] and pyTorch [8].

From both the Royston-Parmar Survival Model (R-PSM) and the Spline Nnet-Survival (SNnet-S), we obtain a functional estimate of the CATE as defined in equation 1, by predicting the measure of effect of interest, $g(\cdot)$, for each combination of the covariates x_1 and by comparing it between the two treatment strategies $z = 0$ and $z = 1$ using the chosen function $f(\cdot)$. The estimation of the CATEs is only the first step of the procedure that leads to the identification of the treatment effect phenotypes. To address the inherent uncertainty in these estimates in subsequent steps and enhance the overall reliability of our findings, we employ resampling techniques to quantify the variability in CATE estimates. For the model-based method, a simulation approach based on parametric bootstrap [9] is applied. For each bootstrap sample, $b = 1, \cdots, B$ the methods consist of:

1. sampling $\boldsymbol{\gamma}^{(b)}$ from the asymptotic Normal distribution of the Maximum Likelihood of the regression parameters,

2. obtaining the corresponding estimates of CATEs $\hat{\tau}_{\mathbf{x}_1}^{(b)}(t)$

On the other hand, for the neural network, non-parametric bootstrap is used. Consequently, for every combination of the covariates $\mathbf{x}_1$ under investigation, we generate a distribution comprising B estimates of CATEs that we will employ in the following step for the identification of the treatment effect phenotypes.

2.3 Identification of treatment effect phenotypes over time

Functional clustering is used to find a grouping structure for $\hat{\tau}_{\mathbf{x}_1}^{(b)}(t)^{[0,w]}$, $b = 1, \cdot, ., B$ for each combination of the covariates $\mathbf{x}_1$ of interest in order to aggregate profiles of individuals in clusters with a common response to the treatment exposure over time. The functional clustering procedure is repeated on each bootstrap sample to take into account the uncertainty in the estimation of the CATEs and improve the stability of the procedure. In FDA, similarly to multivariate data analysis common clustering methods consist of k-means clustering, hierarchical agglomerative clustering or density-based clustering such as the DBSCAN. These, have all been extended to take into account that observations of the group are functional objects [10]. The implementation used in the R package fdacluster [11] has been considered in this work. Specifically, we used functional k-means clustering, choosing the L2 distance between the CATE estimates of the different subject profiles:

$$d_2\{\hat{\tau}_l^{(b)}(w), \hat{\tau}_m^{(b)}(w)\} = \sqrt{\int_0^w \{\hat{\tau}_l^{(b)}(s) - \hat{\tau}_m^{(b)}(s)\}^2 ds} \tag{8}$$
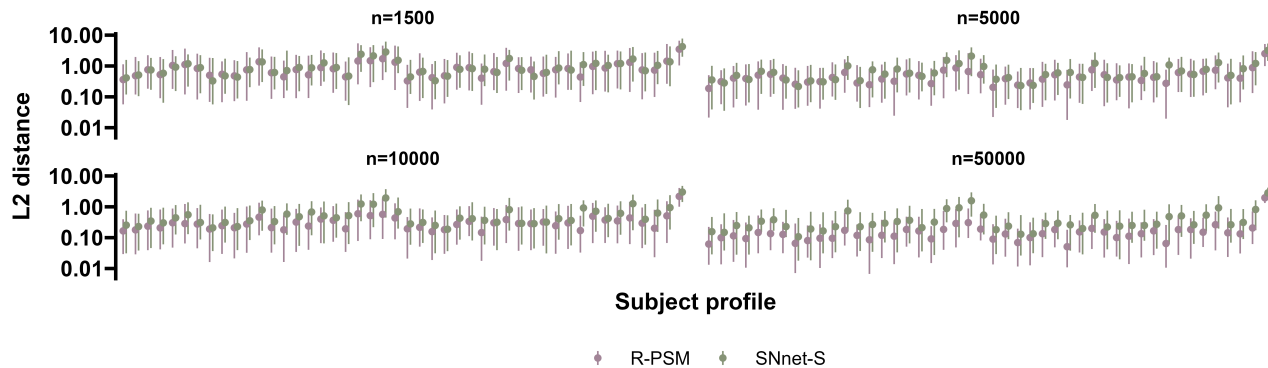
Figure 2: Comparison of the median L2 distance with 95% CI between the estimated cumulative hazard of a correctly specified R-PSM (pink) and SNnet-S (green), relative to the true data generating model.

for $b = 1, \cdots , ..., B$ and each couple of subject profiles l, m. To aggregate the clustering results obtained on the different bootstrap samples, either a simple majority vote or an ensemble method based on consensus clustering can be used. Given a measure of similarity (or agreement) between the results of two clustering, consensus clustering aims at maximizing the average similarity. It is possible to distinguish between medoid consensus clustering [12] and soft consensus clustering [13]. According to medoid consensus, final clustering is sought over the set of the base clusterings. In soft consensus clustering, it is possible to assign objects to several groups with varying degrees of "membership". These clustering ensemble methods are all implemented in R package `clue`[14]. K-means clustering requires choosing the number of clusters. The silhouette values are typically employed for this task. Alternatively, the re-sampling procedure through bootstrap allows us to consider the mean internal agreement between the results of the clustering on the different bootstrap samples to assess the stability of the results with different numbers of clusters. Finally, clusters can be interpreted through the visual inspection of the clustering consensus centroids. These were obtained as the functional median obtained using the Modified Band Depth [15, 16] of the CATEs contained in each final consensus cluster.

3 Simulation Study

3.1 Design

We performed a simulation study to evaluate the performance of the proposed method. In the first part of the simulation, we compare the performance of the SNnet-S with the R-PSM. Secondly, we want to assess the stability of the identification of treatment effect patterns through functional clustering. We considered a vector of 5 binary effect modifiers, and we specified a data-generating model based on the spline-based parametric survival model with time-varying effects and several interactions among the covariates. According to this model, the baseline hazard was specified considering a spline with 1 knot placed at $t = 10$. 100 datasets were simulated under four different sample size scenarios (n=1 500, 5 000, 10 000, 50 000). As a measure of effect in the simulation, we considered the ratio between cumulative hazard.

3.2 Results

The L2 distance between the cumulative hazard of the true data-generating model and the one estimated with either the correctly specified R-PSM or the SNnet-S approach was calculated for each combination of the covariates and binary treatment indicator. The results for the six scenarios are reported in Figure 2. For all sample sizes and all subject profiles, the SNnet-S and the R-PSM reach a similar performance. It is important to note that the performance of R-PSM is conditioned on the fact that we are able to select the correct model.

Established the satisfying performance of the SNnet-S approach, we continued with this approach and performed the functional clustering considering different numbers of clusters $nclust = 2, 3, 4$. 100 was used as the number of bootstrap samples. As previously reported, to ensemble the clustering results obtained in the different bootstrap samples a naive majority vote and consensus clustering methods were considered. Specifically, here we considered three medoid consensus methods and two soft least square consensus methods that differ for the (di-)similarity measure
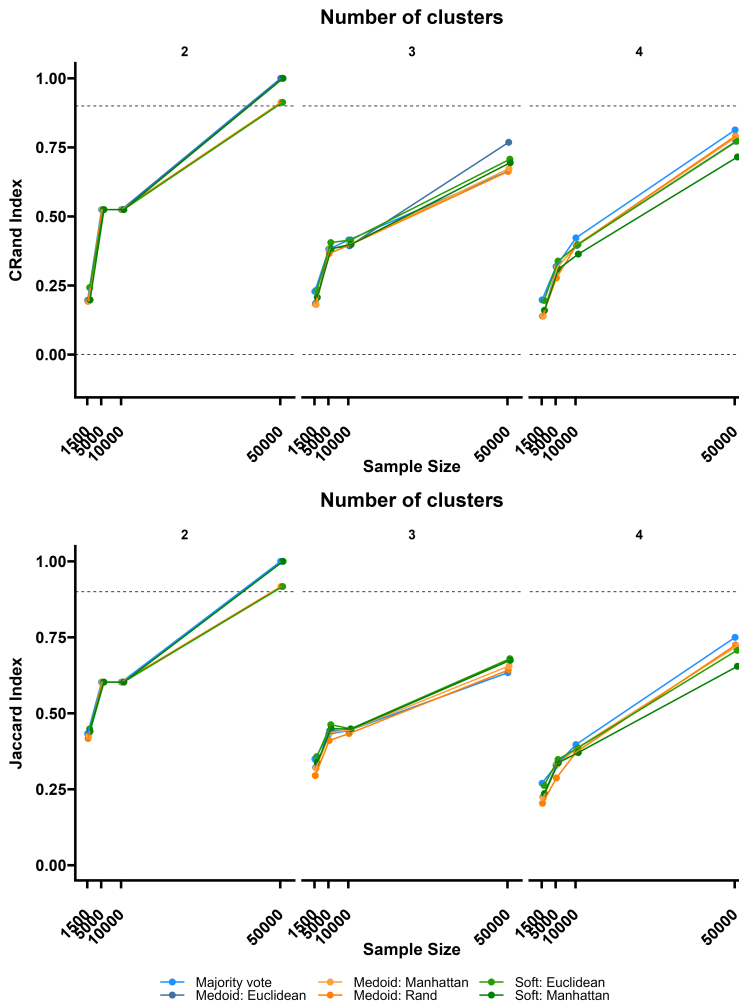
Figure 3: Comparison of the agreement according to the CRand Index (top panel) and the Jaccard Index (bottom panel) according to different bootstrap ensemble methods.

used: Euclidean, Manhattan and Rand are considered for the medoid consensus and, Euclidean and Manhattan are considered for the least square soft consensus.

The results of the clustering obtained on cumulative hazard ratio curves of the data-generating model are considered the "gold standard" and they are compared in terms of agreement across the different ensemble bootstrap methods by means of the Jaccard Index and the Corrected Rand (CRand) Index. The Jaccard index compares the ratio of the numbers of distinct pairs of subject profiles in the same class in both partitions and in at least one partition, respectively, and it can take values from 0 to 1. The CRand Index calculate the rate of distinct pairs of subject profiles both in the same class or both in different classes in both partitions, and it is corrected for agreement by chance. The latter index can take values from -1 to 1 and negative values indicate that agreement is less than what is expected from a random result. In Figure 3 we can observe that the agreement increases with the sample size. Importantly, the results are robust to the choice of the clustering consensus method. The agreement is higher for the two clusters, and this is due to the specific data-generating model used. In particular, considering two clusters, the method achieves an agreement with the "true" clustering above 0.90 according to both indices and all clustering consensus methods on the scenario n=50 000. In Table 1 we report the mean agreement among the clustering results in the 100 bootstrap samples. As expected, a higher internal agreement among clustering samples corresponds to a higher final performance of the method. For two clusters and n=50000, the mean internal agreement is above 0.5.

Table 1: Concordance of clustering results between bootstrap samples

Sample size	Number of clusters	CRand Index	Jaccard
1500	2	0.06	0.38
	3	0.05	0.25
	4	0.04	0.19
5000	2	0.22	0.45
	3	0.19	0.32
	4	0.16	0.25
10000	2	0.30	0.49
	3	0.25	0.36
	4	0.21	0.28
50000	2	0.57	0.65
	3	0.46	0.50
	4	0.41	0.41

4 Application

4.1 Clinical problem & Data

Treatment of Heart Failure relies on several life-saving pharmacological therapies. Among them, MRAs are one of the cornerstones of therapy in heart failure, yet is one that is most often discontinued by cardiologists out of fear of adverse events, e.g. alteration of potassium [17, 18]. However, there is no clear evidence regarding which patients' side effects overcome benefits in terms of the risk of hospitalisation or death. As a consequence, treatment decisions are often made on a subjective basis. The aim of this analysis is to provide the following medical decision: for which profiles of patients discontinuing MRAs during the first year after initiation of the therapy is the safest choice and contrarily, for which type of patient discontinuation leads to higher risks in terms of risk of hospitalization and/or death? According to European guidelines, in deciding whether to discontinue therapy with MRAs cardiologists should consider episodes of hyperkalaemia (high potassium) since starting therapy, age, diabetes, New York Heart Association (N.Y.H.A class), renal function and ejection fraction. Therefore, these variables were assumed to be the effect-modifying variables. In addition, other available variables were considered as possible confounders and were adjusted for in the analysis via propensity score matching.

Data were obtained by the interrogation of the administrative regional health data of Friuli Venezia Giulia Region in the Northern part of Italy, integrated with data derived from the Outpatient and Inpatient Clinic E-chart (Cardionet ®). This integrated database constitutes the Trieste Observatory of Cardiovascular Diseases. Specifically, this was a cohort observational, non-interventional study involving patients living in the Trieste who had a Heart Failure diagnosis between January 2009 and December 2020, had at least one cardiological evaluation, two potassium measurements and, were observed for at least one year after having started therapy with MRAs. For the identification of HF patients, the following steps will be followed. First, a search in the electronic medical records, using appropriate keywords (Heart Failure, Chronic Heart Failure, Systolic Heart Failure, Diastolic Heart Failure) to select patients with HF-related clinical findings. In order to avoid any diagnostic underestimation, data from the medical E-chart were combined with the discharge codes of any previous hospital access (based on the standard nomenclature of the ICD-9 CM) and/or interventional procedures for HF patients (i.e. ICD implantation). Subsequently, prospective cases were manually reviewed by clinicians, to validate the diagnosis of HF using the criteria established in 2016 by the European Cardiology Society. The cohort was followed from the index date, defined as the first date of purchase of MRAs, until an event of cardiovascular hospitalization, death or the end of the follow-up (administrative study closure date, fixed at 31 December 2020). The database has been previously described in the literature [19]. As previously pointed out, all patients had at least one year of observation. The first year will be used as a pre-follow-up period to define the treatment group: persistently treated with MRAs in the first year of therapy vs. therapy with MRAs discontinued during the first year. As a consequence, the analysis will consider the follow-up as the time since the end of the first year of observation.

4.2 Study Cohort

1555 subjects were included in the analysis. The outcome of interest was the first event between an event of hospitalization due to a cardiovascular reason or death. At 2 years of follow-up, the Kaplan-Meier estimate of the event-free probability was equal to 0.66 (95% CI: 0.63-0.68). 64% of the patients are older than 75. 39% present Heart Failure with reduced left-ventricular ejection fraction (<50%) and 14% have severe symptoms (N.Y.H.A III or IV) Moreover, 31%

have diabetes and 47% have Chronic Kidney Disease (CKD). During treatment with MRAs, %11 had hyperkalemic episodes.

4.3 Defining the treatment variable: discontinuation of therapy with MRAs

Using drug prescription data coming from administrative health records, it is possible to define treatment discontinuations using pharmacoepidemiology methods. All prescriptions dispensed to these patients during the pre–follow-up period were identified, and the coverage of each prescription was calculated by dividing the total amount of drug filled in the prescription by the Defined Daily Dose [20]. Subjects were defined as having discontinued treatment with MRAs during the first year of observation if they stopped filling the prescription and a minimum of 90 days had passed from the last day covered by the drug and the start of follow-up. The R package `adhereR`[21] was used to derive the discontinuation measure. 26 % of subjects resulted in having discontinued MRAs in the first year of therapy.

4.4 Controlling for confounding

The propensity score was estimated through logistic regression using variables that were possibly unbalanced in the two exposure groups. Matching was used to obtain a balanced dataset with respect to such variables. Descriptive statistics on the original and matched dataset are reported in Table 2. The matched cohort included 4598 patients. To test for the presence of unbalance, either the Chi-squared, Fisher or t-test were used as appropriate. Interestingly, even before matching, the two groups were already quite balanced.

Table 2: Descriptive Statistics before and after matching.

| | Before Matching | | After Matching | |
	Stand. Effect Size	p-value	Stand. Effect Size	p-value
Sex	-2.90	0.62	-0.35	0.92
Duration of HF	0.41	0.94	4.92	0.17
BMI	-10.65	0.08	-3.40	0.32
Blood Pressure	9.74	0.10	-3.21	0.37
High Heart Rate	-6.21	0.29	5.06	0.16
CPT	15.54	0.01	2.31	0.44
Hypertension	8.00	0.19	-3.49	0.32
Anemia	9.03	0.13	-3.61	0.30
Vasculopathy	-3.01	0.61	-0.28	0.94
BCPO	-1.09	0.85	0.40	0.91
Cancer	0.53	0.93	-0.11	0.98
Ulcer	6.27	0.27	-0.18	0.96
Liver disease	0.27	0.96	-3.61	0.34
Cerebrovascular diseases	3.99	0.50	-0.54	0.88
Dementia	4.84	0.39	-0.27	0.94
Mental Disorders	-1.75	0.77	-4.90	0.23
Number of comorbidities >3	5.88	0.31	-3.56	0.32
ACE/ARBs	-0.37	0.95	-3.13	0.38
Beta-blockers	-6.01	0.31	-2.64	0.47
Diuretics	0.46	0.94	0.09	0.98
Digitalis	-3.37	0.57	5.17	0.13

4.5 Results

The CATEs over a 2-year time frame according to the three effect measures (the hazard ratio, the cumulative hazard ratio and the survival ratio) were estimated using the SNnet-S. 100 bootstrap samples were obtained for each subject profile, and a maximum of four clusters was considered. The average silhouette values and the internal agreement among bootstrap samples favoured two as the number of treatment effect phenotypes. This was consistent across the three measures of effect considered. The average silhouette values and the internal agreement among bootstrap samples favoured two as the number of treatment effect phenotypes. This was consistent across the three measures of effect considered. For this reason, in Figure 4 the medoids of the final clusters for each effect measure for the two-phenotype case are displayed. For both the cumulative hazard ratio and the survival ratio, we see that there is a group (green line) that tends to show a stable protective effect and another that shows a higher risk of events in the first six months.

In fact, in the latter group, this may also be reflected in the hazard, where there is an immediate adverse effect of discontinuation, which tends to decrease rapidly in the first half of the year. However, the overall agreement between the results of the different clusters for all three measures of effect is around 0.24, which means that, due to the sample size, the results are not robust enough to draw any firm conclusions.
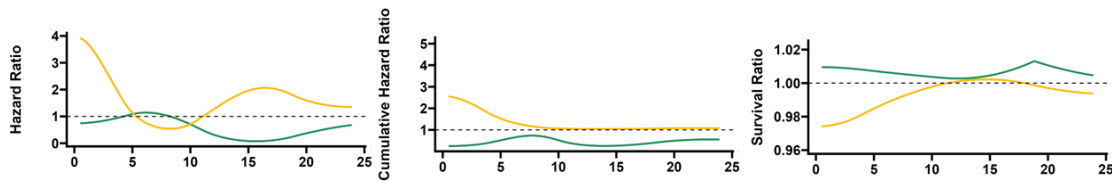

Figure 4: Final clusters medoids for the two-phenotype case.

Nevertheless, as an example, the assignment of the subjects to the phenotypes for the survival ratio and the 2-clusters case is reported in Figure 5.
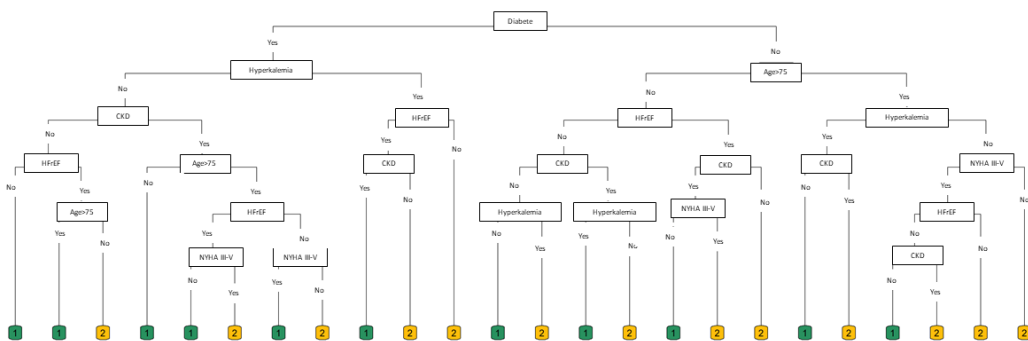

Figure 5: Subjects' assignment to treatment effect phenotypes, considering two as a number of clusters.

5 Discussion and Conclusions

In this research, we have proposed a novel method to characterize which are the different responses to a treatment in a survival setting, taking into account two key aspects: effect heterogeneity and time-varying effects. Most of the methods to inform personalized treatment decisions are built on the assumption that treatment is either "beneficial", "null" or "harmful". However, when considering a time interval as a decision frame of interest, there is a broader range of possibilities. Making decisions without having the full picture of the dynamic response of subjects can lead to unwanted consequences. On the other hand, among the infinite possibilities of responses that can be present in a population, it is essential to summarize to some extent which are the most common treatment-effect phenotypes over time and to which subject profiles they can be attributed.

The aim of the proposed approach is to tackle this issue by a two-stage procedure. First, the conditional average treatment effect over a time span of interest considering relevant effect modifiers was estimated using methods that allow both interactions and non-proportional hazards. Although a spline-based survival model can be useful when there is a strong prior knowledge of the model structure, the approach based on a survival neural network and smoothing of the subject-specific hazard curves is a valid alternative, as shown in the simulation study.

As a second step, an appropriate synthesis that encapsulates the possible responses to the treatment was obtained by clustering the curves representing the conditional treatment effects through functional clustering. While looking at the individual CATEs can be complicated to get a complete picture of the treatment behaviour in the population, clustering the subject-specific treatment effect curves can help to extract a useful summary.

Survival clustering has been already proposed in the literature to discover subpopulations whose survival is regulated by different generative mechanisms [22]. On the other hand, here we are clustering subject-specific treatment effects over time to discover subpopulations that display different treatment effect mechanisms. Nevertheless, in future research, the method could be investigated with the aim of characterizing different risk patterns over time.

We employed the term functional survival analysis to remark that we are treating the CATEs as realizations of functional

9

objects. To our knowledge, this aspect together with the idea of extracting treatment effect phenotypes over time has not been considered before in the literature.

Bootstrapping and consensus clustering were used to take into account the uncertainty in the two-stage procedures that involve the estimation and the clustering. An important aspect concerning the uncertainty that needs to be taken into account is that not all the subject profiles have the same prevalence in the population. Care should be taken in considering the subset of subject profiles for which data is available. In our simulation setting that considered five effect modifiers, the sample size had a strong effect on the accuracy of the clustering results. Indeed, a large number of subjects was needed in order to obtain results in line with the data-generating model. Specific sample size requirements are complex to obtain as they depend on the underlying model. Nevertheless, assessing internal agreement among the results on the different bootstrap samples can serve as a quantification of clustering uncertainty.

In regard to the interpretation of the results, there are two different aspects to consider. The first regards the interpretation of the different phenotypes obtained with respect to the treatment effect over time. This can be done by visualizing the medoids of the different clusters. The second relevant component is which subjects are contained in each of the phenotypes. The mapping between the effect-modifiers space and the phenotypes can be represented by a decision tree, which can be easily used by domain experts to subjects' assignment to a specific phenotype. However, the interpretation of the effect of specific covariates on the CATEs is outside the scope of this work. We have shown a possible application of the proposed approach on data coming from electronic health records and administrative health databases, involving the treatment of Heart Failure patients. In particular, the aim of the study was to assess the response to discontinuation of MRAs during the first year of treatment in the two consequent years. While this issue is very relevant for the treatment of heart failure, the results reported in this work are intended only as exploratory. This application is meant only as an illustrative example of how the proposed method can be used to answer a medical decision question. Analyses on a larger cohort are needed in order to confirm the results so that they can be of use in the clinical management of heart failure patients. An additional challenge in this context concerns the causal effect of interest. We have considered only an intention-to-discontinue effect, as subjects may have re-started treatment with MRAs during the consequent two years. Further directions of this analysis include also studying the per-protocol effect. With respect to the choice of the measure of effect, in survival analysis, Hazard Ratios are traditionally used. However, we have considered the ratio between the cumulative hazards and survival functions, as in the presence, of non-proportional hazards, they allow us to take into account the "accumulation" of risk over time. Although it can be useful to consider both of them, in the medical domain the ratio between the survival curves offers a clearer interpretation. Furthermore, depending on the specific problem, an absolute measure such as the survival difference could be employed instead.

A limitation is the need to choose a number of treatment-effect phenotypes. Even though there are established methods coming from the clustering literature, in this specific context it is also important to confirm the interpretability of the results from a domain-specific point of view. Moreover, the method relies on the choice of a specific clustering algorithm. Although we have considered functional k-mean, it would be possible to use any functional clustering algorithm. Indeed, consensus clustering is independent of the specific clustering method used, and it could in theory also be used to combine the results obtained from different algorithms without the need to choose one algorithm over the other.

References

[1] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MITPress, 2nd edition, 2009.

[2] J. O. Ramsay and B. W. Silverman. Functional Data Analysis. 2005.

[3] Patrick Royston and Mahesh K B Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

[4] Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

[5] Christopher H Jackson. flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of statistical software*, 70, 5 2016.

[6] X.-R. Liu, Y. Pawitan, and M. Clements. Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546, 2018.

[7] Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

[9] Micha Mandel. Simulation-Based Confidence Intervals for Functions With Complicated Derivatives. *http://dx.doi.org/10.1080/00031305.2013.783880*, 67(2):76–81, 5 2013.

[10] Thaddeus Tarpey and Kimberly K J Kinateder. Clustering Functional Data . *Journal of Classification*, 20(1):93–114, 2003.

[11] Aymeric Stamm. fdacluster: Joint Clustering and Alignment of Functional Data, 2023.

[12] Alexander Strehl and Joydeep Ghosh. Relationship-Based Clustering and Visualization for High-Dimensional Data Mining. *INFORMS J. Comput.*, 15:208–230, 2003.

[13] A D Gordon and M Vichi. Fuzzy partition models for fitting a set of partitions. *Psychometrika*, 66(2):229–247, 2001.

[14] Kurt Hornik. A CLUE for CLUster Ensembles. *Journal of Statistical Software*, 14(12 SE - Articles):1–25, 9 2005.

[15] Sara López-Pintado and Juan Romo. On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, 104(486):718–734, 6 2009.

[16] Francesca Ieva, Anna Maria Paganoni, Juan Romo, and Nicholas Tarabelloni. roahd Package: Robust Analysis of High Dimensional Data. *The R Journal*, 11(2):291–307, 2019.

[17] Aldo P Maggioni, Francesco Orso, Silvia Calabria, Elisa Rossi, Elisa Cinconze, Samuele Baldasseroni, Nello Martini, and on behalf of the ARNO Observatory. The real-world evidence of heart failure: findings from 41 413 patients of the ARNO database. *European Journal of Heart Failure*, 18(4):402–410, 4 2016.

[18] Michel Komajda, Stefan D Anker, Martin R Cowie, Gerasimos S Filippatos, Bastian Mengelle, Piotr Ponikowski, and Luigi Tavazzi. Physicians' adherence to guideline-recommended medications in heart failure with reduced ejection fraction: data from the QUALIFY global survey. *European journal of heart failure*, 18(5):514–522, 5 2016.

[19] Annamaria Iorio, Gianfranco Sinagra, and Andrea Di Lenarda. Administrative database, observational research and the Tower of Babel, 6 2019.

[20] WHO Collaborating Centre for Drug Statistics Methodology. *Introduction to drug utilization research*. WHO Library Cataloguing-in-Publication Data, 2003.

[21] Alexandra Lelia Dima and Dan Dediu. Computation of adherence to medication and visualization of medication histories in R with AdhereR: Towards transparent and reproducible use of electronic healthcare data. *PLOS ONE*, 12(4):e0174426, 4 2017.

[22] Laura Manduchi, Ričards Marcinkevičs, Michela C Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C Neidert, Marc Pfister, Bram Stieltjes, and Julia E Vogt. A Deep Variational Approach to Clustering Survival Data, 2022.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

89/2023   Savaré, L.; Ieva, F.; Corrao, G.; Lora, A.
Capturing the variety of clinical pathways in patients with schizophrenic disorders through state sequences analysis

88/2023   Masci, C.; Spreafico, M.; Ieva, F.
Joint modelling of recurrent and terminal events with discretely-distributed non-parametric frailty: application on re-hospitalizations and death in heart failure patients

85/2023   Arnone, E.; De Falco, C.; Formaggia, L.; Meretti, G.; Sangalli, L.M.
Computationally efficient techniques for Spatial Regression with Differential Regularization

86/2023   Ferraccioli, F.; Sangalli, L.M.; Finos, L.
Nonparametric tests for semiparametric regression models

83/2023   Cavinato, L.; Massi, M.C.; Sollini, M.; Kirienko , M.; Ieva, F.
Dual adversarial deconfounding autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data

82/2023   Pozzi, G.; Ciarletta, P.
Geometric control by active mechanics of epithelial gap closure

81/2023   Buchwald, S.; Ciaramella, G; Salomon, J.; Sugny, D.
A SPIRED code for the reconstruction of spin distribution

79/2023   Agosti, A.; Bardin, R.; Ciarletta, P.; Grasselli, M.
A diffuse interface model of tumour evolution under a finite elastic confinement

78/2023   Antonietti, P.F.; Bonizzoni, F.; Corti, M.; Dall'Olio, A.
Discontinuous Galerkin for the heterodimer model of prion dynamics in Parkinson's disease

77/2023   Fumagalli, I.; Corti, M.; Parolini, N.; Antonietti, P. F.
Polytopal discontinuous Galerkin discretization of brain multiphysics flow dynamics