Journal of Data Science 22 (2), 298–313
April 2024

DOI: 10.6339/24-JDS1126
Statistical Data Science

Precision Medicine: Interaction Survival Tree for Recurrent Event Data

Yushan Yang[1], Chamila Perera[2], Philip Miller[3], Xiaogang Su[4], and Lei Liu[3,*]

[1]*Institute for Health and Equity, Medical College of Wisconsin, Milwaukee, WI, USA*
[2]*Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA*
[3]*Division of Biostatistics, Washington University in St. Louis, St. Louis, MO, USA*
[4]*Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX, USA*

Abstract

In randomized controlled trials, individual subjects experiencing recurrent events may display heterogeneous treatment effects. That is, certain subjects might experience beneficial effects, while others might observe negligible improvements or even encounter detrimental effects. To identify subgroups with heterogeneous treatment effects, an interaction survival tree approach is developed in this paper. The Classification and Regression Tree (CART) methodology (Breiman et al., 1984) is inherited to recursively partition the data into subsets that show the greatest interaction with the treatment. The heterogeneity of treatment effects is assessed through Cox's proportional hazards model, with a frailty term to account for the correlation among recurrent events on each subject. A simulation study is conducted for evaluating the performance of the proposed method. Additionally, the method is applied to identify subgroups from a randomized, double-blind, placebo-controlled study for chronic granulomatous disease. R implementation code is publicly available on GitHub at the following URL: https://github.com/xgsu/IT-Frailty.

Keywords *interaction tree; frailty model; subgroup identification*

1 Introduction

In many biomedical studies, a participant might experience recurring instances of the event of interest. These recurrent events can include, for example, recurrent infections, cancer relapses, and repeated hospitalizations. A characteristic shared by these recurrent events is the inherent correlation between the events occurring within the same individual (Amorim and Cai, 2015). Nonetheless, many studies tend to focus solely on the analysis of the initial occurrence of these events (Yang et al., 2017), disregarding the correlation intrinsic to the recurrent events within an individual, which could lead to decreased statistical power and biased estimates. Consequently, it is essential to incorporate the within-individual correlation when modeling these events (Amorim and Cai, 2015).

To fully exploit the recurrent event data, numerous extensions to the original Cox model have been suggested to take subsequent events into account, including Andersen-Gill (Andersen and Gill, 1982), Prentice, Williams, and Peterson (total and gap times) (Prentice et al., 1981), Wei, Lin, and Weissfeld (Wei et al., 1989), and frailty models (e.g., Therneau and Grambsch,

[*] Corresponding author. Email: lei.liu@wustl.edu.

Received August 1, 2023; Accepted March 15, 2024

2000). Other related works include (Pepe and Cai, 1993; Kennedy et al., 2001; Kelly and Lim, 2000)

On the other hand, there is growing importance of detecting heterogeneous treatment effects in the realm of precision medicine. Many tree-based approaches have been proposed (Su et al., 2008, 2009, 2011; Foster et al., 2011; Hao and Zhang, 2014; Kong et al., 2017) for subgroup identification, a key aspect of comparative analysis that focuses on evaluating the impact of treatment on responses. The aim is to understand the heterogeneity of the treatment effect across different subpopulations. Techniques such as the Interaction Tree (IT) procedure utilize recursive partitioning to conduct subgroup identification and can autonomously discover several objectively defined subgroups (Su et al., 2009). Within some of these subgroups, individual subjects may experience beneficial effects, while subjects in other subgroups may observe negligible or even detrimental effects. However, it is noted that these interaction tree methodologies have yet to be adapted for the analysis of recurrent event data.

Our study is motivated by a randomized, double-blind, placebo-controlled study (The International Chronic Granulomatous Disease Cooperative Study Group, 1991) on chronic granulomatous disease (CGD), an uncommon inherited disorder typically initiating early in life and potentially leading to childhood mortality. This trial was designed to assessed the overall efficacy of interferon gamma in treating serious infections among patients with CGD. The primary endpoint is the recurrence of infections tracked for approximately one year. The results of the trial established a significant efficacy of interferon compared to placebo, both in preventing the first serious infection and in reducing the frequency of recurrent serious infections. The study also reported that subgroups with age less than 10, X-linked disease, or autosomal recessive disease, showed more beneficial effects. However, the study only considered one variable at a time and did not provide the rationale for selecting subgroups, e.g., an age cutoff of 10. Our method aims to seek objectively and optimally defined subgroups and explore differential treatment effects on recurrent events by fully leveraging all the characteristics in the dataset. Understanding the potential heterogeneity of treatment effects can provide valuable insights for tailoring treatment strategies to individual patients, thereby optimizing the treatment of CGD.

The primary goal of this paper is to extend the interaction tree approach to recurrent event data, enabling the identification of subgroups characterized by heterogeneous treatment effects. While various approaches are available to model recurrent event data, our focus is on the frailty Cox model (Clayton, 1978), where a log-normal frailty term is employed to account for the correlation among recurrent events in the same individual. We implement the proposed method with R (R Core Team, 2024), a widely used statistical software.

The rest of the paper is organized as follows. In Section 2, we introduce the model and method proposed in this study. Section 3 presents the results of the simulation studies. In Section 4, we apply our method to the data from the randomized clinical trial on chronic granulomatous disease. Finally, Section 5 concludes the paper, summarizing our findings and discussing the implications of this work.

2 Methods

Recurrent event data, often encountered in medical and epidemiological studies, capture the occurrences of repeated events over time, while accounting for the influence of covariates on the event processes. Consider a typical study involving n subjects, each of whom may experience multiple recurrent events. The resultant dataset can be presented as $\mathcal{D} = \{(Y_{ik}, \delta_{ik}, \mathrm{trt}_i, \boldsymbol{x}_{ik}) :$

$i = 1, \ldots, n; k = 1, \ldots, n_i\}$, where subject i has n_i recurrent events. $Y_{ik} = \min(V_{ik}, C_i)$ is the k-th observed recurrent time before censoring time C_i, and V_{ik} denotes the k-th true recurrence time from baseline. $\delta_{ik} = I(V_{ik} \leqslant C_i)$ is an indicator of recurrent or censoring status, where $I(\cdot)$ is the indicator function. trt_i is a binary treatment indicator with 1 for treated and 0 for untreated, $\boldsymbol{x}_{ik} \in \mathbb{R}^p$ is the associated covariate vector.

In subgroup identification, it is interesting to see how baseline covariates contribute to the heterogeneity of treatment effects. Therefore, we assume all covariates are measured at baseline only, i.e., $\boldsymbol{x}_{ik} \equiv \boldsymbol{x}_i$.

2.1 Frailty Model

Numerous approaches are available for modeling recurrent event time data, with three major types being conditional models, marginal models, and frailty models. The outcome under consideration for modeling can be either the time since study entry, also known as the total time (TT), or the gap time (GT) since the previous event. For an in-depth exploration, readers are referred to Cook and Lawless (2007), Therneau and Grambsch (2000), and Amorim and Cai (2015). Among these approaches, the frailty model extends the Cox's proportional hazards model by introducing a random effect to account for dependence among the recurrent event times (Clayton, 1978; Kelly and Lim, 2000). This random effect represents additional risk or frailty for individual subjects, capturing unmeasured characteristics that cannot be explained by observed covariates alone. The frailty model assumes that the correlated event times become independent when conditioning on the covariates and random effects.

In this study, we consider frailty models of the form as follows

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i + \eta_i), \tag{1}$$

where $\lambda_i(t)$ is the hazard (intensity) function for the recurrent time at time t (measured from baseline) of subject i, $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}$ is the vector of coefficients for the covariate $\boldsymbol{x}_i$, and η_i is the frailty or random effect terms for subject i. The frailty term η_i is often assumed to follow either a Gaussian or log-gamma distribution. As recommended by the R package **coxme** (Therneau, 2024), the Gaussian distribution aligns naturally with (generalized) linear mixed models, enabling the incorporation of more complex variance-covariance structures for random effects (Ripatti and Palmgren, 2000). In this study, we adhere to **coxme** and assume a normal distribution $\mathcal{N}(0, \sigma^2)$ for η_i.

2.2 Interaction Tree (IT)

Interaction Tree (IT) (Su et al., 2008; Hou et al., 2015) is a tree-based method specifically designed for subgroup identification. Its strength lies in its ability to handle non-linear covariate effects and complex interactions. IT recursively partitions the data by identifying the best split that captures the highest heterogeneity in treatment effects, resulting in a hierarchical tree structure. Consequently, subjects in the terminal nodes with the highest beneficial treatment effects are regarded as the most responsive to the treatment. The graphical tree representation enhances the interpretability of IT analysis. With its capability to discover meaningful subgroups characterized by distinct treatment responses, IT becomes a valuable tool for improving treatment effectiveness and advancing precision medicine.

2.3 IT for Recurrent Event Data

Following a CART convention, IT analysis consists of three major steps: growing a large tree, pruning, and tree size selection. Details will be given below.

2.3.1 Growing a Large Tree

In the first stage of IT analysis, a large preliminary tree is developed by continuously dividing data into two subsets that show the greatest variation in treatment effects. For subject i, denote the variables of interest as $\{X_{ij}, j = 1, \ldots, J\}$. The heterogeneity of treatment effects is assessed through the following frailty model with an interaction term:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 \mathrm{trt}_i + \beta_2 Z_i + \beta_3 Z_i \times \mathrm{trt}_i + \eta_i), \tag{2}$$

where the frailty term η_i follows $\mathcal{N}(0, \sigma^2)$ and $Z_i = I(X_{ij} \leqslant a)$ is the indicator variable associated with the binary split s for an ordinal or continuous covariate X_{ij}. If X_{ij} is nominal with levels $L = \{l_1, \ldots, l_r\}$, then $Z_i = I\left(X_{ij} \in A\right)$ for a subset $A \subset L$ is considered.

To evaluate the split s, it is natural to consider testing hypotheses: $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$. For Cox frailty models, three standard tests are available: the likelihood ratio test (LRT), the score test, and the Wald test, all being asymptotically equivalent. The score test is often deemed more computationally efficient in tree modeling. However, this is generally not the case for interaction trees since the main effect term of the split indicator Z_i is necessarily included in the model. While all these choices are included in our implementation, we have primarily used the Wald statistic as the splitting statistic, denoted as $G(s)$, to assess the split s. In addition, it is important to note that a splitting statistic can be similarly developed from any of conditional or marginal models (Prentice et al., 1981; Andersen and Gill, 1982; Wei et al., 1989). Nonetheless, our exposition is mainly focused on the frailty model-based approach.

For a given split s, the splitting statistic $G(s)$ follows the chi-square $\chi^2(1)$ distribution with one degree of freedom under the null hypothesis H_0. At node h, all permissible splits are considered. The greedy search step constitutes the primary source of computational burden in the entire interaction tree analysis. For a continuous X_j (hereafter we use X_j instead of X_{ij} for simplicity of notation), permissible splits involve all its distinct observed values that satisfy additional stopping criteria. For a categorical X_j, examining all the subsets of its levels can be computationally intensive. One remedy is to sort these levels according to the estimated treatment effect within each level and then treat X_j as an ordinal covariate. The split $s^\star$ with the maximum splitting statistic, denoted as $G(h)$,

$$G(h) = G(s^\star) = \max_s \ G(s),$$

which is associated with the most significant interaction with treatment, is selected as the best split to bisect node h into two child nodes.

A large initial tree T_0 is obtained by recursively applying the same procedure on child nodes until some loosely defined stopping rules are met. Common stopping criteria in interaction trees include maximum tree size, maximum tree depth, and minimum node size. The minimum node size refers to the minimum number of subjects in either the treated or untreated group, separately defined for node h and its subsequent child nodes. For recurrent event times, additional stopping rules for the minimum number of events should also be enforced.

2.3.2 Pruning

The large initial tree T_0 contains both the important true model structure as well as superfluous splits. A subtree of T_0 will be selected as the final model. Nevertheless, T_0 has a massive number of subtrees, which make it unfeasible to consider all possible subtrees. Therefore, branches of T_0 are iteratively trimmed, one at a time, until we reach the null tree model, which consists solely of the root node. This results in a much smaller subset of subtrees for the subsequent assessment and selection.

Variants of pruning have been proposed. Because IT splits by maximizing the between-node difference, the maximized score statistic $G(s)$ is an innate measure of goodness-of-split for each internal node. Consequently, a split-complexity pruning algorithm for trees grown by goodness-of-split (LeBlanc and Crowley, 1993) is naturally well-suited for IT.

The algorithm is based on the following split-complexity measure:

$$G_\alpha(T) = G(T) + \alpha|T'| = \sum_{h \in T'} G(h) + \alpha|T'|, \tag{3}$$

where T' denotes all the internal nodes of tree T and $|\cdot|$ means cardinality; the goodness-of-split measure $G(T) = \sum_{h \in T'} G(h)$ represents the amount of heterogeneity in treatment effects represented by tree T. The total number of internal nodes of tree T, $|T'|$, is a measure of tree complexity; and the complexity parameter $\alpha > 0$ penalizes $G(T)$ for each added split. With fixed α, the larger $G_\alpha(T)$, the better tree model T.

As α increases from 0, there will be an internal node h of T_0 that first becomes ineffective, in the sense that the subtree after trimming the branch rooting from node h is more effective than the subtree including that branch. This link h is then identified as the weakest link and its associated branch is trimmed, yielding a subtree $T_1 \preceq T_0$. Next, keep increasing α to find the weakest link of T_1 and trim it. Repeating this procedure leads to a limited number of subtrees, $T_0 \succeq T_1 \succeq T_2 \succeq \dots \succeq T_M$, which forms a nested sequence of optimally pruned subtrees.

2.3.3 Tree Size Selection

From the sequence, one subtree will be chosen to serve as the final tree model. The same split-complexity measure $G_\alpha(T)$ serves as the criterion for model assessment comparison. To this end, the goodness-of-split measure $G(T)$ can be overoptimistic (biased upwards) due to the adaptive nature of the greedy search used in constructing the tree. As a result, a more 'honest' estimate of $G(T)$ is needed. To do so, a test sample approach can be employed if the sample size is sufficiently large. In cases with moderate or small sample sizes, a bootstrap method for bias correction can be applied. In this method, the bias of $G(T)$ is first estimated through repeated resampling. Subsequently, this estimated bias is applied to adjust $G(T)$, resulting in a new estimate of $G(T)$ with reduced bias. One is referred to LeBlanc and Crowley (1993) or Su et al. (2009) for details.

Furthermore, the complexity parameter α is fixed for tree size selection purposes. A value within the range $2 \leqslant \alpha \leqslant 4$ is suggested in LeBlanc and Crowley (1993), where $\alpha = 2$ aligns with the Akaike information criterion (AIC) (Akaike, 1973) and $\alpha = 4$ corresponds roughly to the 0.05 significance level on the $\chi^2(1)$ curve. Another viable choice is the logarithm of the effective sample size in spirit of BIC (Schwarz, 1978). It is worth noting that the 'effective sample size' in the context of recurrent events lacks a well-defined definition, with various options such as the number of subjects (n), the number of recurrent or censoring events in total (N), the number of recurrent events ($\sum_{i=1}^{n} \sum_{k=1}^{n_i} \delta_{ik}$), and the number of subjects with at least one event

$\left(\sum_{i=1}^{n} I\left(\sum_{k=1}^{n_i} \delta_{ik} > 0\right)\right)$. On the other hand, these choices typically do not differ much from each other on the logarithmic scale unless in extreme scenarios. We have chosen to experiment with $\log(n)$ and $\log(N)$ in simulation studies presented in Section 3. The final interaction tree $T^\star$ is selected as the one with maximum $G_\alpha(T)$.

3 Simulation Studies

In this section, simulation studies are designed to assess the performance of the proposed interaction survival tree method in subgroup analysis of recurrent event data. Through these simulation studies, we aim to evaluate the accuracy, robustness, and effectiveness of our method in subgroup identification and variable selection.

Each data set includes four covariates X_1 to X_4, which are generated independently from Uniform(0, 1). In addition, a binary treatment variable 'trt', is generated from a Bernoulli distribution with $p = 0.5$. Setting $Z_1 = I\{X_1 \leqslant 0.5\}$ and $Z_2 = I\{X_2 \leqslant 0.5\}$, we then generate the recurrent event times from the following Cox frailty model:

$$\lambda_i(t) = \omega_i\, \lambda_0(t) \exp\left(\beta_1 \mathrm{trt}_i + \beta_2 Z_{i1} + \beta_3 Z_{i2} + \beta_4 Z_{i1} Z_{i2} \mathrm{trt}_i\right), \tag{4}$$

with $(\beta_1, \beta_2, \beta_3)^T = (-1, 1, 1)^T$, where the multiplicative frailties $\omega_i = \exp(\eta_i)$ are generated either from log-normal or a Gamma distribution. In either case, ω_i has mean 1 and variance $\theta = 1$. When the log-normal distribution is used for ω_i, this amounts to simulating η_i from $\mathcal{N}(0, \sigma = 0.6935)$. When ω_i follows the Gamma distribution, its density function is given by

$$f(\omega) = \frac{\omega^{1/\theta - 1} e^{-\omega/\theta}}{\Gamma(1/\theta)\theta^{1/\theta}},$$

for $\omega > 0$ and $\theta > 0$. This enables us to assess the robustness of our method when faced with potential misspecification of the frailty distribution. In addition, the coefficient β_4 corresponds to the moderating effect of covariates on the treatment effect. We will manipulate β_4 to explore different levels of signal strength.

For the baseline hazard function $\lambda_0(t)$, we examine two choices: the exponential baseline hazard with $\lambda_0(t) = 0.25$, and the Weibull baseline hazard with $\lambda_0(t) = 2t$. To generate recurrent event times with the Weibull baseline hazard, we essentially compute $V_{i(k+1)} = \sqrt{V_{ik}^2 - \log U_{ik}/\tau_i^2}$ for the i-th subject, where V_{ik} is the k-th recurrent event time with $V_{i0} = 0$, $U_{ik} \sim$ Uniform(0, 1), and $\tau_i^2 = \omega_i \exp\left(\beta_1 \mathrm{trt}_i + \beta_2 Z_{i1} + \beta_3 Z_{i2} + \beta_4 Z_{i1} Z_{i2} \mathrm{trt}_i\right)$. Finally, we simulate a censoring time C_i from Uniform (0, 5) for subject i so that $Y_{ik} = \min(V_{ik}, C_i)$. By simulating different scenarios and varying key parameters, we can thoroughly examine the performance of the interaction survival tree approach under various conditions and gain insights into its strengths and limitations.

Given the moderately small sample size in the colorectal cancer study data, we specifically focus on investigating the bootstrap method for tree size selection. Due to the intensive computational nature of this approach, we are constrained to examine a limited number of scenarios. Each model setting is examined for 100 simulation runs with sample size $n = 300$. For each generated dataset, the number of bootstrap samples is set to 20. Six choices, $\{1, 2, 3, 4, \log(n), \log(N)\}$, are used for the complexity parameter α in determining the final tree model. In addition, the minimum number of events in each arm in each terminal node is 3 with minimum node size 20. Performance metrics include variable selection, the number of null final trees (# T_0), and summaries of final tree sizes. In terms of variable selection, we count the number of final trees that contain X_1, X_2, X_1 & X_2, and X_1 or X_2. To summarize the final tree sizes, we report the mean and SD.

Table 1: Simulation study I: the null case with $\beta_4 = 0$ in Model (4) with log-normal frailties.

		# Trees with Variables				#	Tree Sizes	
$\lambda_0(t)$	α	X_1	X_2	X_1 & X_2	X_1 or X_2	T_0	Mean	SD
Exponential	1	46	52	22	76	14	3.12	0.977
	2	31	34	13	52	40	2.42	1.232
	3	16	21	6	31	63	1.85	1.175
	4	8	14	4	18	78	1.45	0.892
	$\log(n)$	2	4	2	4	95	1.09	0.429
	$\log(N)$	0	0	0	0	100	1.00	0.000
Weibull	1	51	50	24	77	8	3.31	0.838
	2	38	38	15	61	29	2.67	1.164
	3	22	23	7	38	53	1.94	1.108
	4	9	15	0	24	71	1.50	0.870
	$\log(n)$	1	3	0	4	95	1.07	0.326
	$\log(N)$	0	0	0	0	100	1.00	0.000

3.1 Study I: The Null Case

We begin by examining the null case where no treatment-by-covariate interaction is present. To achieve this, we set $\beta_4 = 0$ in Model (4) with log-normal frailties. Consequently, the covariates X_1 and X_2 are only involved with additive effects. This setup enables us to thoroughly investigate the Type I error or false signal issue that may arise in subgroup analysis, which is of utmost importance for all subgroup analysis approaches (Sleight, 2000).

Table 1 provides a summary of the results from 100 simulation runs under the null case. It is evident that the choice of α plays a central role in preventing false positive or Type I errors of subgroup identification. When $\alpha \in \{1, 2\}$, the likelihood of obtaining the true null tree structure is only 40% or less. With higher values of α, false signals are better controlled. While this process differs from traditional statistical hypothesis testing, one may wish to maintain a small probability of committing a Type I error, such as 0.05 or 0.10. In this regard, only BIC-typed penalties, i.e., $\log(n)$ or $\log(N)$, appear to meet this requirement effectively. This result generally aligns well with those reported in Su et al. (2009) and Su et al. (2011). In subgroup identification, it is crucial to differentiate between prognostic and predictive covariates. With a relatively larger penalty, the interaction tree approach proves effective in discerning between additive effects and interactions (Loh et al., 2019). Despite the inclusion of X_1 and X_2 in the true model, their effects remain additive, thus seldom selected by the interaction tree (IT). These conclusions hold true for both scenarios with constant or linear baseline hazard functions. This underscores the efficacy of our method in precisely identifying and capturing significant treatment-covariate interactions while avoiding false detection of non-existent interactions in the data.

3.2 Study II: A Tree Model

Next we investigate the case when the true model is a tree-structured model. To achieve this, we set $\beta_4 = 3$ in Model (4) with log-normal frailties. This results in an underlying model that

Table 2: Simulation study II: Tree Model (4) with $\beta_4 = 3$ and log-normal frailties. The true tree model has 3 terminal nodes induced by splits $X_1 \leqslant 0.5$ and $X_2 \leqslant 0.5$.

		# Trees with Variables				#	Tree Sizes	
$\lambda_0(t)$	α	X_1	X_2	X_1 & X_2	X_1 or X_2	T_0	Mean	SD
Exponential	1	93	90	83	100	0	4.50	1.030
	2	93	90	83	100	0	3.94	1.003
	3	92	89	81	100	0	3.56	0.957
	4	91	88	79	100	0	3.20	0.865
	$\log(n)$	91	88	79	100	0	2.98	0.619
	$\log(N)$	91	88	79	100	0	2.88	0.518
Weibull	1	94	95	89	100	0	4.41	0.944
	2	93	94	87	100	0	3.86	0.910
	3	91	93	84	100	0	3.44	0.833
	4	91	93	84	100	0	3.21	0.671
	$\log(n)$	91	93	84	100	0	3.00	0.569
	$\log(N)$	91	92	83	100	0	2.88	0.477

can be represented by a tree structure with three terminal nodes induced by two splits $X_1 \leqslant 0.5$ and $X_2 \leqslant 0.5$. Our primary interest is to examine whether the interaction tree (IT) method can accurately identify the true tree model and correctly select both X_1 and X_2 as the splitting variables.

Table 2 provides the summarized results from 100 simulation runs. The IT method demonstrates a notable ability to recover the true structure in the majority of cases. Regardless of the choice of penalty parameters, a reasonably averaged tree size is achieved. However, when $\alpha \in \{1, 2\}$ is small, some overfitting may occur. Conversely, employing a larger α results in a reduction in the final tree size, mitigating the overfitting effect. As anticipated, both X_1 and X_2 emerge as predominant splitting variables in the final trees, indicating their consistent importance in shaping the tree's structure. The choice of α has a minimal impact on variable selection results. Despite the added complexity introduced by the Weibull baseline, the IT method performs quite comparably to the case with a constant baseline hazard. This demonstrates the efficiency and robustness of IT, which can be attributed to the semi-parametric nature of the Cox frailty model.

3.3 Study III: Misspecified Frailty Distribution

In this investigation, we assess the sensitivity of our method to potential misspecification of the frailty distribution. For this purpose, we generate data from Model (4) with gamma frailty, but utilize the log-normal frailty, as implemented in the R package **coxme** (Therneau, 2024), to develop the interaction trees. We experiment with both the null model and the tree model, and the outcomes are summarized in Table 3.

Compared to the results in Tables 1 and 2, it is evident that our method demonstrates remarkable robustness against frailty distribution misspecification overall, with only a slight drop in performance. When data are generated from the null model, a larger penalty, especially the BIC-typed ones, well prevents false positive errors. In such instances, the null tree model is

Table 3: Simulation study III: misspecified frailty distribution. The null case has $\beta_4 = 0$ and the tree model has $\beta_4 = 3$ in Model (4) with gamma frailty.

			# Trees with Variables				#	Tree Sizes	
Model	$\lambda_0(t)$	α	X_1	X_2	X_1 & X_2	X_1 or X_2	T_0	Mean	SD
Null	Exponential	1	47	49	22	74	12	3.20	0.943
		2	32	34	12	54	36	2.48	1.193
		3	19	19	6	32	60	1.88	1.166
		4	6	10	2	14	81	1.41	0.877
		$\log(n)$	2	6	0	8	88	1.20	0.586
		$\log(N)$	1	2	0	3	94	1.08	0.339
	Weibull	1	54	42	21	75	4	3.36	0.704
		2	39	28	12	55	28	2.67	1.146
		3	24	16	6	34	51	2.03	1.150
		4	14	6	3	17	73	1.50	0.893
		$\log(n)$	6	0	0	6	91	1.14	0.472
		$\log(N)$	2	0	0	2	96	1.05	0.261
Tree	Exponential	1	91	95	86	100	0	4.53	0.904
		2	91	95	86	100	0	4.00	0.943
		3	89	95	84	100	0	3.51	0.937
		4	89	95	84	100	0	3.18	0.702
		$\log(n)$	89	95	84	100	0	2.98	0.550
		$\log(N)$	89	95	84	100	0	2.91	0.452
	Weibull	1	90	95	85	100	0	3.92	0.971
		2	87	94	81	100	0	3.61	0.994
		3	87	92	79	100	0	3.31	0.873
		4	87	91	78	100	0	3.13	0.747
		$\log(n)$	86	89	75	100	0	2.87	0.544
		$\log(N)$	85	88	73	100	0	2.77	0.489

predominantly identified across 100 simulation runs. This outcome holds true irrespective of the baseline hazard chosen. When data are generated from the tree structured model, our method performs effectively in identifying the true tree structure in the majority of cases. As anticipated, slight overfitting occurs when the penalty parameter α is too small. However, a larger α results in a more parsimonious final tree, with the average tree size closely approximating its expected value of 3. In terms of variable selection, either X_1 or X_2 is consistently chosen to split the root node in the final tree. Moreover, over 70% of the final trees have both X_1 and X_2 as splitting variables, effectively capturing the true tree structure induced by the splits $X_1 \leqslant 0.5$ and $X_2 \leqslant 0.5$.

3.4 Study IV: Weaker Signal with Unbalanced Cut

In this study, we change the definition of Z_1 to $Z_1 = I\{X_1 \leqslant 0.2\}$ and apply a weaker signal for the interaction term by letting $\beta_4 = 1$. This allows us to investigate how the proposed tree

Table 4: Simulation study IV: weaker signal with unbalanced cut. The tree model has $\beta_4 = 1$ in Model (4) and splits induced by $X_1 \leqslant 0.2$ and $X_2 \leqslant 0.5$.

		# Trees with Variables				#	Tree Sizes	
$\lambda_0(t)$	α	X_1	X_2	X_1 & X_2	X_1 or X_2	T_0	Mean	SD
Exponential	1	90	78	70	98	2	5.71	1.028
	2	81	62	56	87	13	4.48	1.714
	3	63	44	37	70	28	3.18	1.743
	4	49	30	20	59	41	2.22	1.338
	$\log(n)$	30	15	8	37	63	1.59	0.911
	$\log(N)$	21	13	6	28	72	1.41	0.767
Weibull	1	88	63	56	95	1	4.88	0.998
	2	75	47	39	83	12	3.94	1.455
	3	59	32	23	68	28	2.87	1.488
	4	42	18	13	47	47	2.04	1.214
	$\log(n)$	25	7	5	27	69	1.49	0.823
	$\log(N)$	22	6	3	25	72	1.38	0.678

method works under a different scenario with an unbalanced cutoff and a weaker signal. Table 4 presents the summarized results out of 100 simulation runs.

It can be seen that the complexity stemming from unbalanced cuts and weak signal strength poses challenges in identifying the true tree structure. When using BIC-typed penalties, the weak signal can scarcely be discerned due to the stringent penalty imposed. Conversely, employing a smaller penalty, such as $\alpha = 2$ or 3, proves advantageous. In these instances, a final tree of average size around 3, induced by either X_1 or X_2, consistently emerges. Therefore, the penalty parameter plays a crucial role in striking a balance between preventing Type I errors and effectively identifying signals within the data. In practical data analysis, it is advisable to experiment with various choices of α and examine the resulting tree models.

Overall, these results affirm the capability of our method to correctly identify significant treatment-covariate interactions. The frailty IT also shows strong robustness and flexibility in handling various scenarios arising from frailty distributions, baseline hazards, unbalanced cuts, and signal strengths. These simulation studies provide valuable guidance for the application of our method in real-world settings and contribute to its potential use in clinical and research settings for analyzing recurrent event data.

4 Application to Chronic Granulomatous Disease Data

To illustrate our proposed method, we consider the CGD dataset, which consists of 128 patients enrolled in a randomized, double-blind, placebo-controlled study (The International Chronic Granulomatous Disease Cooperative Study Group, 1991). The primary outcome, the times of infection recurrence, is subject to censorship. The dataset also includes nine baseline characteristics measured at study entry: enrolling center (`center`, 13 centers in total), sex (`sex`), age in years (`age`), height in cm (`height`), weight in kg (`weight`), mode of inheritance (`inherit`), use of steroids (`steroids`), use of prophylactic antibiotics (`propylac`), categorization of the centers into 4 groups (`hos.cat`).

The mean age of the entire dataset is 14.64, with a median of 12 and a maximum value of 44. Consequently, the patients in this dataset are predominantly children or adolescents. Height and weight are pivotal variables in the context of CGD, and it is documented that patients with CGD typically exhibit reduced heights and weights (Pietro Bortoletto et al., 2015). Body Mass Index (BMI) provides a more meaningful interpretation than considering height and weight separately. However, it is inappropriate to directly compare the BMI of adults with that of children, as the interpretation of BMI differs significantly between these two groups due to differences in body composition, growth patterns, and developmental stages. Therefore, our analysis focuses solely on patients aged 19 or under, encompassing a total of 90 subjects. Instead of the standard BMI measurement, we compute the BMI-for-age *z*-scores, which is a standardized measure indicating how a child's BMI compares to the distribution of BMIs among children of the same age and sex in a reference population (Must and Anderson, 2006; WHO Multicentre Growth Reference Study Group, 2006). The BMI-for-age z-score represents the number of standard deviations (SD) from the mean. For example, a *z*-score of +1 indicates that the child's BMI is one standard deviation above the mean in the reference population (Martinez-Millana et al., 2018).

Table 5 furnishes a summary of baseline characteristics for patients aged 19 or under. The dataset comprises 90 patients, among whom 32 experienced at least one recurrent event, with

Table 5: Characteristics for patients in chronic granulomatous disease data.

Characteristic	Overall (N=90)	Placebo (N=43)	Treatment (interferon gamma) (N=47)
Age at baseline			
Mean(SD)	9.24 (5.10)	9.16 (5.58)	9.32 (4.68)
Median(Range)	8.50 (1.00, 19.00)	8.00 (1.00, 19.00)	9.00 (1.00, 19.00)
Sex			
Male	77 (86%)	35 (81%)	42 (89%)
Female	13 (14%)	8 (19%)	5 (11%)
Inherit			
X-linked	66 (73%)	28 (65%)	38 (80%)
Autosomal	24 (27%)	15 (35%)	9 (20%)
Steroids			
Yes	1 (1%)	0 (0%)	1 (2%)
No	89 (99%)	43 (100%)	46 (98%)
Proplylac			
Yes	82 (91%)	38 (89%)	44 (96%)
No	8 (9%)	5 (11%)	3 (4%)
Hos.cat			
US:other	47 (52%)	24 (56%)	23 (49%)
Europe:other	14 (16%)	9 (21%)	5 (11%)
US:NIH	18 (20%)	6 (14%)	12 (25%)
Europe:Amsterdam	11 (12%)	4 (9%)	7 (15%)
BMI z-score			
Mean(SD)	-0.07 (1.30)	-0.24 (1.20)	0.07 (1.38)
Median(Range)	-0.02 (-3.38, 3.76)	-0.10 (-3.30, 1.83)	0.06 (-3.38, 3.76)

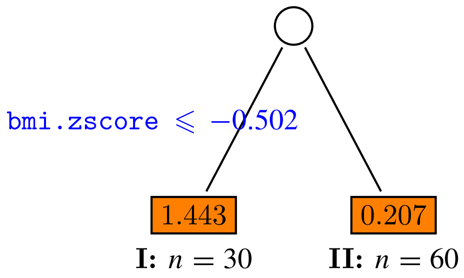

Figure 1: The final interaction tree structure for the chronic granulomatous disease (CGD) recurrent event data. The treatment effect, quantified by hazard ratio, is specified within each terminal node. The sample size (number of subjects) is also indicated beneath each terminal node.

one patient encountering as many as seven recurrent events. This results in a total of 60 recurrent events. Censoring occurred at the last observation.

We apply the IT method to the dataset with the following default constraints: a minimum of 2 subjects with at least one recurrent event in either child node of a split, a minimum node size for further split set at 10, and a maximum tree depth of 6. The final tree is determined with the bootstrap method, using $B = 30$ bootstrap samples. The IT approach involves growing a large initial tree to capture essential structures. At the same time, a sufficient number of observations, particularly uncensored event times, at each split and in each child node are required to estimate the frailty model and avoid numerical difficulties. These specified constraints are designed to address these considerations.

The final tree structure, selected with $\alpha = \log(n)$, where n is the number of subjects, is depicted in Figure 1. This choice of the final tree ensures a strict control of false signals, as indicated by our earlier simulation studies. The tree consists of two terminal nodes (labeled as I, II) determined by the split: `bmi.zscore` $\leqslant -0.5$. A BMI z-score of -0.5 indicates that the patient's BMI is 0.5 standard deviations below the mean BMI for their age and sex group in the reference population. Group I comprises patients with a relatively lower body mass than Group II.

To assess the stability of the final tree size, we conducted a sensitivity analysis by varying the minimum node size $\{5, 10, 15\}$ and the minimum number of subjects with at least one recurrent event in either child node of a split $\{1, 2, 3\}$. The penalty parameter α is set to $\log(n)$ to mitigate the risk of false signals. The final tree sizes under different constraints are depicted in Figure 2. It can be seen that when the minimum number of subjects with at least one recurrent event in either child node of a split is set to 2, the final tree size remains stable across varying minimum node sizes. Nevertheless, when the minimum number of subjects with at least one recurrent event in each arm in each terminal node is set to 3 or greater, only the root node is present. This is mainly attributed to the dataset's relatively small size, resulting in a significant influence of this constraint on the bootstrap samples. Adjusting these constraints provides increased flexibility for the IT method, aiding in the identification of final tree models that are more interpretable and robust.

Table 6 provides a brief summary of the two subgroups identified by the final IT tree, along with the entire data. Among the child or adolescent patients with CGD under study, the overall

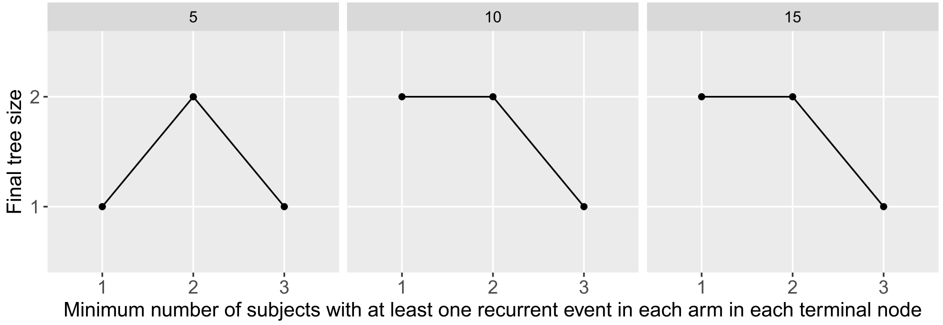

Figure 2: Final tree sizes by varying the minimum node size {5, 10, 15} and the minimum number of subjects with at least one recurrent event in either child node of a split {1, 2, 3}.

Table 6: Summary of subgroups for the chronic granulomatous disease data.

			# Patients (n)		
	Trt Effect	P-value	Placebo	Treatment	Total
Group I	1.443	0.546	15	15	30
Group II	0.207	<0.001	28	32	60
All Group	0.361	0.003	43	47	90

treatment effect, as indicated by the estimated hazard ratio obtained from fitting a frailty Cox model with **coxme**, is 0.36 (p-value < 0.01). Group I, consisting of 30 patients with a BMI z-score equal to or less than -0.5, exhibits a treatment effect of 1.443 (p-value $= 0.55$). The remaining subjects form Group II, comprising 60 patients with a treatment effect of 0.207 (p-value < 0.01). The substantial heterogeneity in the treatment effects indicates that while interferon gamma was found to be overall significantly efficacious among children or adolescent CGD patients, it might not be helpful or could even have a detrimental effect for those who are relatively underweight (with a BMI z-score equal to or less than -0.5). The findings underscore the importance of considering BMI as a crucial factor when prescribing interferon gamma treatment.

As a cautionary note, extra care should be exercised in interpreting the p-values presented in Table 6, as they should not be treated as those stemming from traditional hypothesis testing. Given the highly adaptive nature of the tree method, inherent over-optimism may arise. Therefore, validation with future independently collected new data is essential to strengthen the robustness and generalizability of our findings. This will instill greater confidence in the identified subgroups and enable a more accurate assessment of their corresponding treatment effects.

Given that the IT method and the subsequent *post-hoc* analyses are all rooted in frailty proportional hazards (PH) models, it is prudent to examine the assumption of the proportionality of treatment effects. Following Therneau and Grambsch (2000), we initially fitted the frailty PH model with treatment included, obtaining empirical Bayes frailty estimates. Subsequently, we fitted a regular Cox proportional hazards model, including the logarithm of the empirical Bayes frailty estimates as an offset term. Finally, the Schoenfeld (1982) test was conducted to assess the proportional hazards assumption, resulting in a p-value of 0.22. Therefore, we conclude that the proportional hazards assumption is not violated in our analysis.

5 Discussion

In this paper, we extend the interaction tree (IT) method to identify subgroups with recurrent event data. Subgroup analysis essentially involves treatment-by-covariates interactions while tree models are well-suited for handling complex interactions. By introducing tree methods into subgroup analysis, IT inherits the merits of ordinary tree models such as robustness to distributional assumptions due to its nonparametric nature, highly interpretable results, and invariance to monotone transformations on covariates. As a result, IT serves as a powerful tool for post-hoc subgroup analysis.

Our method can be extended in several directions. First, we can consider scenarios where the observation of recurrent events is terminated by informative dropouts or dependent death events. To address this, we can apply the IT to shared frailty models of both recurrent and terminal events (Liu et al., 2004). Second, alternative modeling approaches for recurrent event data, such as models for between-event or gap times, can be integrated with interaction trees for subgroup analysis. Although our main emphasis has been on frailty models specifically using lognormal frailty in this paper, incorporating gamma frailty can be seamlessly achieved using the R function `coxph()`. Moreover, other more flexible frailty distributions can be adopted, leveraging methods like the probability integral transformation (Liu and Yu, 2008) or likelihood reformulation (Nelson et al., 2006). Third, in the current method, we have assumed proportionality. However, for future work, we are interested in exploring non-proportional hazards models for recurrent events, such as the transformation models (Zeng and Lin, 2007). These extensions would further enhance the versatility and applicability of our method in analyzing recurrent event data. Finally, we have implemented our method on clinical trial data employing double-blind randomization. Adapting this method for use in observational studies represents an intriguing direction for future research.

Funding

This research is partly supported by NIH grants R21 AG084054 and UL1 TR002345.

References

Akaike H (1973). Information theory and an extension of the maximum likelihood principle. In: *Selected Papers of Hirotugu Akaike*, 199–213. Springer New York, New York, NY.

Amorim LD, Cai J (2015). Modelling recurrent events: A tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1): 324–333. https://doi.org/10.1093/ije/dyu222

Andersen PK, Gill RD (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10: 1100–1120.

Breiman L, Friedman J, Olshen R, Stone C (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Clayton D (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1): 141–151. https://doi.org/10.1093/biomet/65.1.141

Cook R, Lawless J (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York, NY.

Foster JC, Taylor JM, Ruberg SJ (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24): 2867–2880. https://doi.org/10.1002/sim.4322

Hao N, Zhang HH (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507): 1285–1301. https://doi.org/10.1080/01621459.2014.881741

Hou J, Seneviratne C, Su X, Taylor J, Johnson B, Wang XQ, et al. (2015). Subgroup identification in personalized treatment of alcohol dependence. *Alcoholism, Clinical and Experimental Research*, 39(7): 1253–1259. https://doi.org/10.1111/acer.12759

Kelly PJ, Lim LLY (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine*, 19(1): 13–33. https://doi.org/10.1002/(SICI)1097-0258(20000115)19:1<13::AID-SIM279>3.0.CO;2-5

Kennedy BS, Kasl SV, Vaccarino V (2001). Repeated hospitalizations and self-rated health among the elderly: A multivariate failure time analysis. *American Journal of Epidemiology*, 153(3): 232–241. https://doi.org/10.1093/aje/153.3.232

Kong Y, Li D, Fan Y, Lv J (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics*, 45(2): 897–922. https://doi.org/10.1214/16-AOS1474

LeBlanc M, Crowley J (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422): 457–467. https://doi.org/10.1080/01621459.1993.10476296

Liu L, Wolfe RA, Huang X (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3): 747–756. https://doi.org/10.1111/j.0006-341X.2004.00225.x

Liu L, Yu Z (2008). A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, 27(16): 3105–3124. https://doi.org/10.1002/sim.3153

Loh WY, Cao L, Zhou P (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery*, 9(e1326): 1–21.

Martinez-Millana A, Hulst JM, Boon M, Witters P, Fernandez-Llatas C, Asseiceira I, et al. (2018). Optimisation of children z-score calculation based on new statistical techniques. *PLoS ONE*, 13(12): e0208362. https://doi.org/10.1371/journal.pone.0208362

Must A, Anderson S (2006). Body mass index in children and adolescents: Considerations for population-based applications. *International Journal of Obesity*, 30(4): 590–594. https://doi.org/10.1038/sj.ijo.0803300

Nelson KP, Lipsitz SR, Fitzmaurice GM, Ibrahim J, Parzen M, Strawderman R (2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with non-normal random effects. *Journal of Computational and Graphical Statistics*, 15(1): 39–57. https://doi.org/10.1198/106186006X96854

Pepe MS, Cai J (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88(423): 811–820. https://doi.org/10.1080/01621459.1993.10476346

Pietro Bortoletto P, Lyman K, Camacho A, Fricchione M, Khanolkar A, Katz B (2015). Chronic granulomatous disease. *The Pediatric Infectious Disease Journal*, 34: 1110–1114. https://doi.org/10.1097/INF.0000000000000840

Prentice RL, Williams BJ, Peterson AV (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2): 373–379. https://doi.org/10.1093/biomet/68.2.373

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripatti S, Palmgren J (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4): 1016–1022. https://doi.org/10.1111/j.0006-341X.2000.01016.x

Schoenfeld D (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1): 239–241. https://doi.org/10.1093/biomet/69.1.239

Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464. https://doi.org/10.1214/aos/1176344136

Sleight P (2000). Debate: Subgroup analyses in clinical trials: Fun to look at – but don't believe them! *Current Controlled Trials in Cardiovascular Medicine*, 1(1): 25–27. https://doi.org/10.1186/CVM-1-1-025

Su X, Meneses K, McNees P, Johnson WO (2011). Interaction trees: Exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 60(3): 457–474. https://doi.org/10.1111/j.1467-9876.2010.00754.x

Su X, Tsai CL, Wang H, Nickerson DM, Li B (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2).

Su X, Zhou T, Yan X, Fan J, Yang S (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1), Article 2.

The International Chronic Granulomatous Disease Cooperative Study Group (1991). A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. *The New England Journal of Medicine*, 324(8): 509–516. https://doi.org/10.1056/NEJM199102213240801

Therneau TM (2024). ***coxme**: Mixed Effects Cox Models*. R package version 2.2-20.

Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Wei LJ, Lin DY Weissfeld L (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408): 1065–1073. https://doi.org/10.1080/01621459.1989.10478873

WHO Multicentre Growth Reference Study Group (2006). WHO child growth standards based on length/height, weight and age. *Acta Pdæiatrica. Supplement*, 450: 76.

Yang W, Jepson C, Xie D, Roy JA, Shou H, Hsu JY, et al. (2017). Statistical methods for recurrent event analysis in cohort studies of ckd. *Clinical Journal of the American Society of Nephrology*, 12(12): 2066. https://doi.org/10.2215/CJN.0000000000000302

Zeng D, Lin D (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102(477): 167–180. https://doi.org/10.1198/016214506000001239