Received: 27 August 2019 | Revised: 14 December 2020 | Accepted: 30 December 2020

DOI: 10.1002/sim.9132

RESEARCH ARTICLE

Statistics in Medicine WILEY

Lasso estimation of hierarchical interactions for analyzing heterogeneity of treatment effect

Yu Du[1] | **Huan Chen**[2] | **Ravi Varadhan**[2,3]

[1]Department of Biometrics, Eli Lilly and Company, Indianapolis, Indiana

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

[3]Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland

Correspondence
Ravi Varadhan, Division of Biostatistics and Bioinformatics, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA.
Email: ravi.varadhan@jhu.edu

Individuals differ in how they respond to a given treatment. In an effort to predict the treatment response and analyze the heterogeneity of treatment effect, we propose a general modeling framework by identifying treatment-covariate interactions honoring a hierarchical condition. We construct a single-step l_1 norm penalty procedure that maintains the hierarchical structure of interactions in the sense that a treatment-covariate interaction term is included in the model only when either the covariate or both the covariate and treatment have nonzero main effects. We developed a constrained Lasso approach with two parameterization schemes that enforce the hierarchical interaction restriction differently. We solved the resulting constrained optimization problem using a spectral projected gradient method. We compared our methods to the unstructured Lasso using simulation studies including a scenario that violates the hierarchical condition (misspecified model). The simulations showed that our methods yielded more parsimonious models and outperformed the unstructured Lasso for correctly identifying nonzero treatment-covariate interactions. The superior performance of our methods are also corroborated by an application to a large randomized clinical trial data investigating a drug for treating congestive heart failure (N = 2569). Our methods provide a well-suited approach for doing secondary analysis in clinical trials to analyze heterogeneous treatment effects and to identify predictive biomarkers.

KEYWORDS

heterogeneity of treatment effect, hierarchical interaction, Lasso, treatment-covariate interaction

1 | INTRODUCTION

Individuals differ in how they respond to a given treatment. Characterizing the predictors of treatment effect becomes ever increasingly important in the investigation of heterogeneity in treatment effect. This heterogeneity in treatment response is well recognized in clinical practice. Hence any summary from a clinical trial, such as the overall treatment effect, is not directly relevant for treating individual patients.[1-3] A new treatment might only benefit a subpopulation of the patients with certain characteristics, while it shows no benefit to others. Therefore, finding predictive factors, the covariates predictive of treatment response, becomes an important step that could provide insights for personalized medicine. In Epidemiology such variables are called effect modifiers, namely, the treatment effect varies across individuals for different values of these variables. Predictive factors are often derived from prognostic factors, the variables imposing an impact on the outcome in the absence of the treatment. Often, there are numerous candidate prognostic factors, and the task is to find which ones are predictive of treatment response.

There is a robust literature addressing the estimation of heterogeneity of treatment effect including, for example, References 4-15. More recently, there is an emerging literature on machine learning methods for estimating heterogeneous treatment effects, including References 16-19, and so on. A common strategy for studying the treatment response heterogeneity in clinical trial is subgroup analysis, which explores how treatment effect varies across subgroups. However, traditional methods like one-variable-at-a-time subgroup analysis[20] ignore the joint effect of the covariates on treatment effect. Furthermore, they may fail to identify significant treatment-covariate interactions when the number of variables is fairly large, which is often the case in clinical trial studies. Reference 21 considers multivariate subgroups. But, it is still in the subgroup analysis context since the covariates are categorical. Our setting is broader in the sense that we considered both continuous and categorical covariates. Furthermore, subgroup analysis problems are typically set up as hypothesis testing problems. Whereas, our interest lies in modeling the heterogeneity of treatment effect, that is, identifying covariates that are predictors of treatment response. Another approach would be to prespecify the prognostic covariates and fit a model with all treatment-covariate interactions. This is termed as unstructured interaction model.[22] However, this approach has two important limitations. The variables need to be prespecified, and the model may include interactions that are not actually present.

Reference 22 proposed a parsimonious approach, extending the work of Reference 23, to use proportional interactions model to investigate treatment response heterogeneity in a randomized controlled clinical trial. This approach jointly considers the effect modification of various variables, however, it still has difficulty in dealing with fairly large number of candidate effect modifiers. Another limitation is the potential for model misspecification where the underlying treatment-covariate interactions are not proportional to the main effects.

We propose a general method in this article that overcomes the limitations of Reference 22 to assess heterogeneity of treatment response in clinical trials setting. Our work differs in two aspects: (1) We relax the "proportional" constraint, allowing more flexibility in estimating treatment-covariate interactions; (2) the methods are able to automatically screen a large number of potential effect modifiers, with desirable properties of a good trade-off balance of false-positives and false-negatives to correctly capture the significant interactions.

Our work is also inspired by the work of Reference 24, which modifies Lasso[25] to estimate a sparse interaction model, considering the complete list of two-way interactions. One important constraint employed by Reference 24 is the interaction hierarchy condition that an interaction term can only be included in the model when one or both of the associated variables are retained in the model. Our proposed methods make use of this principle. This principle is considered very practical, as Reference 26 once stressed that "Large component main effects are more likely to lead to appreciable interactions than small components. In addition, the interactions corresponding to larger main effects may be in some sense of more practical importance."

Our work differs from Reference 24 in two ways. First, we focus on treatment-covariate interactions in the context of randomized clinical trials, whereas Reference 24 examines all two-way interactions in a context different than evaluating treatment effect heterogeneity. Second, we propose two different ways to incorporate hierarchical constraints. Thus, our work provides a new and potentially useful framework for identifying predictive variables for analyzing heterogeneous treatment response.

The modeling of treatment-covariate interactions have also been considered by References 27,28, among others. We approached the modeling differently in that we employed the hierarchical structure of interactions that help to strike a good balance between false negatives and false positives in uncovering the true effect modifiers.

Given a large set of p covariates from a randomized clinical trial, we select and estimate a subset of candidate effect modifiers that are predictive of treatment response. Our modeling approach is fully parametric and provides a clear interpretation of how individual baseline characteristics affect treatment response. The primary outcome we consider in this article is time-to-event, although our proposed methods can be easily adapted to other types of endpoints, for example, continuous or binary. In Section 2, we introduce the notation, interaction hierarchy restriction, and the parametric modeling assumption. We constructed a single-step l_1 norm penalty procedure that maintains the hierarchical structure of interactions. We studied two parameterization schemes with different constraints honoring hierarchical condition. The formulated constrained optimization problems were solved by using spectral projected gradient (SPG) method.

We examined the performance of our proposed methods in simulation studies in Section 3 including a scenario where the interaction hierarchy restriction was violated and compared with unstructured Lasso. In Section 4, we applied our methods to the Studies of Left Ventricular Dysfunction Treatment (SOLVD-T) trial, a two-arm placebo-controlled randomized clinical trial investigating the efficacy of enalapril, the angiotensin-converting-enzyme inhibitor, to reduce the hazard of death or hospitalization among patients with chronic heart failure.[29] The results of additional simulations based on the real trial are also presented. We provide a discussion in Section 5 of the proposed methods, and discuss future extensions of this work.

2 | METHOD

2.1 | Notations, assumptions, and the assumed model

In this article, we target the time-to-event (TTE) outcome, where we use T to denote the event time, and C, the censoring time. The noninformative censoring is assumed throughout. The observed outcome is represented by a vector (X, Δ), $X = \min(T, C)$, and $\Delta = \mathrm{I}(T \leq C)$, where $\mathrm{I}(T \leq C)$ is the indicator variable taking value 1 if $T \leq C$. Suppose we are in the context of a clinical trial with two arms in parallel. Let A be the treatment indicator, where $A = 1$ means assignment to the treatment arm while $A = 0$ means assignment to the control arm. Let Z be the vector of p candidate effect modifiers, such that $Z = (Z_1, Z_2, \dots , Z_p)'$. Thus, the complete observed data for subject i is denoted by the vector $D_i = (X_i, \Delta_i, A_i, Z_i)$, $i = 1, 2, \dots , n$, assuming that there are n observations in total. We assume the following multivariable Cox proportional hazards model[30] to relate the TTE outcome distribution to the subject's treatment assignment, p candidate effect modifiers as well as p candidate interaction terms:

$$\lambda(t|A_i, Z_i) = \lambda_0(t) \exp\left(\beta_A A_i + \beta_Z' Z_i + \gamma' A_i Z_i\right), \tag{1}$$

where $\lambda_0(t)$ represents baseline hazard at time t for any subject i randomly drawn from the overall population, $\lambda(t|A_i, Z_i)$ gives the hazard function at time t conditioned on the subject i treatment assignment and his/her candidate effect modifiers. In this Cox model (1), β_A, the main effect of treatment, is interpreted as the log hazard ratio comparing treatment to control, while $\beta_Z = (\beta_{Z_1}, \beta_{Z_2}, \dots , \beta_{Z_p})'$ is a p element vector denoting the prognostic effect of Z, and $\gamma = (\gamma_1, \gamma_2, \dots , \gamma_p)'$ is also a p element vector showing the treatment-covariate interaction effects. Any nonzero $\gamma_j, j = 1, 2, \dots , p$ in the vector γ identifies an effect modification from variable Z_j on treatment A in a sense that treatment effect varies across different values of Z_j. This causes a heterogeneity in treatment effect across subjects with varied values of Z_j in the population. The advantage of assuming a fully parametric model incorporating pairwise interactions between treatment and covariates is that we can provide a clear interpretation of how individual baseline characteristics affect treatment response. For example, for the variable Z_j, $\gamma_j > 0$ indicates a heterogeneity in a direction that the treatment becomes less efficacious for higher values of Z_j while $\gamma_j < 0$ implies a stronger treatment response when Z_j increases. Let us use θ to represent the vector of all the parameters, $\theta = (\beta_A, \beta_Z', \gamma')'$.

2.2 | Parameterization schemes and optimization problems

Interaction hierarchy restriction states that an interaction term should be included in the model only when the corresponding main effects are present in the model. This type of restriction has also been known as "heredity" and "marginality"—see, for example, References 31,32 and 33 among others. We adapt the definition to the clinical setting in the context of a clinical trial with two arms in parallel. The goal is to capture treatment-covariate interactions and to inform the variables predictive of treatment response. Therefore, we define a hierarchical structure for treatment-covariate interaction such that for $j = 1, 2, \dots , p$

$$\gamma_j \neq 0 \Rightarrow \beta_{Z_j} \neq 0. \tag{2}$$

We make direct use of this hierarchy in the parameterization schemes that will follow. Traditionally, the parameter vector θ is estimated by minimizing over θ without any constraints the negative log partial likelihood (namely, maximizing the partial likelihood), as defined by $l(\theta)$ such that

$$l(\theta) = -\sum_{i:\Delta=1} \log \frac{\lambda(X_i|A_i, Z_i)}{\sum_{j:X_j \geq X_i} \lambda(X_i|A_j, Z_j)} \tag{3}$$

$$= -\sum_{i:\Delta=1} \left(\beta_A A_i + {\beta_Z}' Z_i + \gamma' A_i Z_i - \log \sum_{j:X_j \geq X_i} \exp\left(\beta_A A_j + {\beta_Z}' Z_j + \gamma' A_j Z_j\right) \right). \tag{4}$$

It is often the case that only a handful of estimates out of many have nonzero effects in the model, corresponding to a sparse structure. To address the sparsity, Lasso is a widely applied technique proposed by Reference 25 that employs model selection and estimation at the same time. This is accomplished by imposing an l_1 norm penalty on the parameter vector and solving a constrained convex optimization problem:

$$\begin{aligned} &\underset{\theta}{\text{Minimize}} && l(\theta) \\ &\text{subject to} && \|\theta_1\| \leq \lambda, \end{aligned} \tag{5}$$

where λ acts as the penalty parameter, balancing the trade-off between fitting to the training data and the enforcement of a sparse coefficients structure. The less the value of λ is, the more sparse the estimates of the parameter vector θ will be. We proposed two parameterization schemes shown below to extend Lasso[25] that guarantee to produce models satisfying interaction hierarchy restriction.

- Parameterization Scheme 1 (PS1)

$$\begin{aligned} &\underset{\theta}{\text{Minimize}} && l(\theta) \\ &\text{subject to} && \gamma_j = \beta_{Z_j} * \zeta_j, \quad \text{for } j = 1, 2, \dots, p, \\ & && \|\beta_A\|_1 + \|\beta_Z\|_1 + \|\zeta\|_1 \leq \lambda. \end{aligned} \tag{6}$$

Note that PS1 reparameterizes the coefficient γ_j as a product of the corresponding main effect β_{Z_j} and a new created variable ζ_j. The product structure from PS1 guarantees a nonzero interaction $\hat{\gamma}_j$ if and only if $\hat{\beta}_j \neq 0$. The parameters defined in PS1 are estimable.

- Parameterization Scheme 2 (PS2)

$$\begin{aligned} &\underset{\theta}{\text{Minimize}} && l(\theta) \\ &\text{subject to} && |\gamma_j| \leq |\beta_{Z_j}|, \quad \text{for } j = 1, 2, \dots, p, \\ & && \|\gamma\|_1 \leq |\beta_A|, \\ & && \|\theta\|_1 \leq \lambda. \end{aligned} \tag{7}$$

PS2 is adapted from the added constraint in Reference 24 to fit the clinical setting where only treatment-covariate interactions are considered. The absolute value inequalities on the vector γ and each of its parameters enforce the interaction hierarchy restriction. A nonzero $\gamma_j, j = 1, 2, \dots, p$ translates to a nonzero β_{Z_j}.

2.3 | Algorithm to solve the optimization problems

Despite the enforcement of the interaction hierarchy restriction, the optimization problems with PS2 (7) listed in Section 2.2 are not convex, thus difficult to solve. The nonconvexity comes from the condition $|\gamma_j| \leq |\beta_{Z_j}|, j = 1, \dots, p$. This constraint defines a nonconvex region of the parameter space. We thus implement simple convex relaxation of the problems by replacing each parameter with two of its components such that for each $j = 1, 2, \dots, p$,

$$\gamma_j = \gamma_j^+ - \gamma_j^-, \tag{8}$$

$$\beta_{Z_j} = \beta_{Z_j}^+ - \beta_{Z_j}^-, \tag{9}$$

$$\beta_A = \beta_A^+ - \beta_A^-, \tag{10}$$

TABLE 1 Algorithm for spectral projected gradient method

Input:
$l(\theta)$, the negative log partial likelihood of survival data;
$\nabla l(\theta)$, the gradient function of $l(\theta)$;
λ_0, the initial value of the spectral step length;
$\mathbb{P}_\Omega()$, the projection function into the convex set Ω of the constraints;
$\theta_0 \in \Omega$, the initial value of θ;
ϵ, the value of tolerance.
Goal:
find the minimizer $\hat{\theta}$ of $l(\theta)$ subject to $\hat{\theta} \in \Omega$.
Algorithm:
at the kth iteration, $k \geq 1$,
while not convergent, for example, $\|\mathbb{P}_\Omega(\theta_k - \nabla l(\theta_k)) - \theta_k\|_\infty > \epsilon$,
do:
compute the search direction $d_k = \mathbb{P}_\Omega(\theta_k - \lambda_k \nabla l(\theta_k)) - \theta_k$,
compute the step length α_k,
compute $\theta_{k+1} = \theta_k + \alpha_k d_k$,
compute spectral step length λ_{k+1}.

$$|\gamma_j| = \gamma_j^+ + \gamma_j^-, \tag{11}$$

$$|\beta_{Z_j}| = \beta_{Z_j}^+ + \beta_{Z_j}^-, \tag{12}$$

$$|\beta_A| = \beta_A^+ + \beta_A^-, \tag{13}$$

where $\gamma_j^+ \geq 0, \gamma_j^- \geq 0, \beta_{Z_j}^+ \geq 0, \beta_{Z_j}^- \geq 0, \beta_A^+ \geq 0, \beta_A^- \geq 0$. After the decomposition procedure, this original problem becomes a convex problem with affine (linear) constraints. No absolute values are included in the constraints. Traditionally, the positive part α^+ and negative part α^{-1} of a variable α are defined to be

$$\alpha^+ = \max(\alpha, 0), \quad \alpha^- = -\min(\alpha, 0) \tag{14}$$

Hence it is inherited from this definition that $\alpha^+\alpha^- = 0$. But it is important to note that $\gamma_j^+, \beta_{Z_j}^+, \beta_A^+, \gamma_j^-, \beta_{Z_j}^-, \beta_A^-$ defined here do not require the conditions $\gamma_j^+\gamma_j^- = 0$, $\beta_{Z_j}^+\beta_{Z_j}^- = 0$, and $\beta_A^+\beta_A^- = 0$ because those conditions make the problem nonconvex and intractable to solve. In addition, when the conditions on $\gamma_j^+\gamma_j^- = 0$, $\beta_{Z_j}^+\beta_{Z_j}^- = 0$, and $\beta_A^+\beta_A^- = 0$ are not specified, it is plausible that $\gamma_j^+\gamma_j^- > 0$, $\beta_{Z_j}^+\beta_{Z_j}^- > 0$, and $\beta_A^+\beta_A^- > 0$. The constraints consequently become less restrictive because, for example, it is possible that the solution to such optimization problems returns both $\hat{\beta}_{Z_j}^+ > 0$ and $\hat{\beta}_{Z_j}^- > 0$ for the estimate of β_{Z_j}. Therefore, conditions like $|\gamma_j| \leq |\beta_{Z_j}| = \beta_{Z_j}^+ + \beta_{Z_j}^-$ could have a larger upper bound. This notwithstanding, such a convex relaxation still guarantees interaction hierarchy, as proved in Reference 24. For PS1, it is an advantage that the added constraints do not alter the convexity of the problem, although a potential disadvantage is that it increases the nonlinearity. However, this is not a major issue computationally since the objective function, the negative partial log-likelihood, is already a nonlinear function.

We applied the SPG method proposed by Reference 34 to solve the optimization problems in Section 2.2. As its name suggests, the SPG method incorporates the spectral gradient scheme[35] to greatly improve the effectiveness of the gradient projection method (Bertsekas[36] and references therein). As pointed out in Reference 37, the SPG method has been applied in wide areas of statistics, becoming an ideal tool for large-scale convex constrained optimization problems. The general SPG algorithm used in this article is listed in Table 1, where $\theta = (\beta_A, \beta_Z', \gamma')'$ except in PS1 where $\theta = (\beta_A, \beta_Z', \zeta')'$ due to the reparameterization of γ.

The computation of the step length α_k and the spectral step length λ_k are given in details in Reference 37. An implementation of the SPG method is readily available in R,[38] provided by the *BB* package written by Reference 39. The function *spg*() in the *BB* package provides three different options for spectral step lengths: (1) the step length used in Reference 34; (2) the step length proposed in Reference 40; (3) the step length proposed by Reference 41. As recommended by Reference 39, we used the function *spg*() with the third choice. Using the function *spg*() requires us to provide the computation for $l(\theta)$, the negative log partial likelihood of survival data; $\nabla l(\theta)$, the gradient function of $l(\theta)$; and $\mathbb{P}_{\Omega}()$, the function to project any arbitrary point into the feasible convex set Ω of the constraints. Once these arguments are provided, the function returns the minimizer $\hat{\theta}$ of the function $l(\theta)$, and any nonzero $\hat{\gamma}_j$ reveals a significant treatment-covariate interaction.

Alternative methods are available to solve these convex constrained optimization problems, for example, by using a sequential quadratic programming algorithm written by Reference 42, among others. We explored that method in the simulation study and it produced very similar results and performance to SPG method. It remains an area for future work to compare different optimization algorithms and find the most efficient method. However, this article focuses on proposing two parameterization schemes that extend Lasso and honor the interaction hierarchy restriction in an effort to identify the treatment-covariate interactions out of many candidate effect modifiers.

3 | SIMULATION

3.1 | Simulation setup

In this section, we conduct several simulations to study the performance of the interaction hierarchy model in the context of two parameterizations listed in Section 2.2. One thousand clinical trials are simulated such that each trial assigns $n = 200$ patients to treatment and control arm with 1:1 randomization ratio.

Each trial comes with $p = 50$ potential prognostic variables. Thus, 50 candidate treatment-covariate interactions are there. We let 25 prognostic variables out of 50 have significant impact on the outcome (ie, with nonzero coefficients). We set five treatment-covariate interaction coefficients to be nonzero. Two scenarios are utilized as the generating distributions. For Scenario (A), to generate $\gamma_j = \beta_{Z_j} \cdot \zeta_j$, β_A is set to be 1, β_{Z_j} is generated from the distribution $N(0, 1)$ and ζ_j is generated from the uniform distribution $U[-1, 1]$. For Scenario (B), both β_{Z_j} and γ_j are independently generated from $N(0, 1)$. The censoring rate is fixed at 0.1.

The primary outcome is TTE, a survival endpoint, generated by an exponential distribution assuming a proportional hazards model (1). The censoring time is generated from an independent exponential distribution. The simulation study aims to evaluate the performance of these methods for identifying nonzero treatment-covariate interactions in the presence of (i) hierarchical interactions (Scenario (A)) and (ii) a violation of such interaction hierarchy (Scenario (B)). Therefore, we mainly consider two scenarios for the underlying data generating distribution:

(A) Hierarchical interaction is enforced,

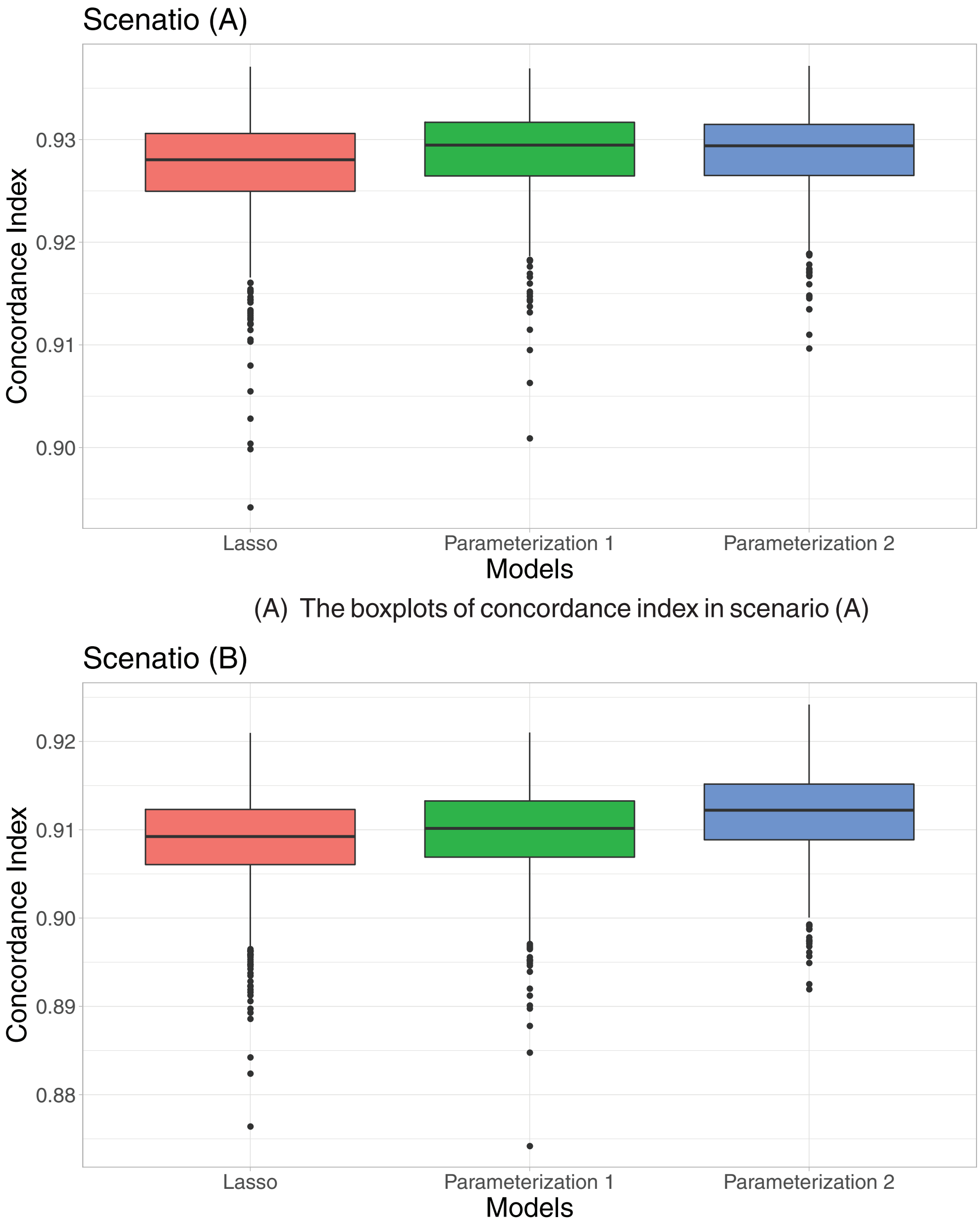
$$\gamma_j \neq 0 \quad \Rightarrow \quad \beta_{Z_j} \neq 0,$$

(B) A violation of interaction hierarchy restriction (model misspecification).

$$\beta_{Z_j} = 0 \quad \not\Rightarrow \quad \gamma_j = 0.$$

3.2 | Simulation evaluation and results

We evaluate the performance of the proposed hierarchical formulations under the two aforementioned situations: (i) when the fundamental assumption about the interaction hierarchy is satisfied (Scenario (A)), and (ii) when the interaction hierarchy is violated, so that the impact of model misspecification can be assessed (Scenario (B)). Lasso, unmodified, serves as a basis of comparison since this is a simple and straightforward approach most commonly used by statisticians for parameter regularization and identification of a sparse structure of the coefficients. The evaluation of performance is twofold: (1) prediction performance and (2) the ability of models to correctly recover the nonzero treatment-covariate interactions. In the assessment of prediction performance, we use the concordance index, as shown by



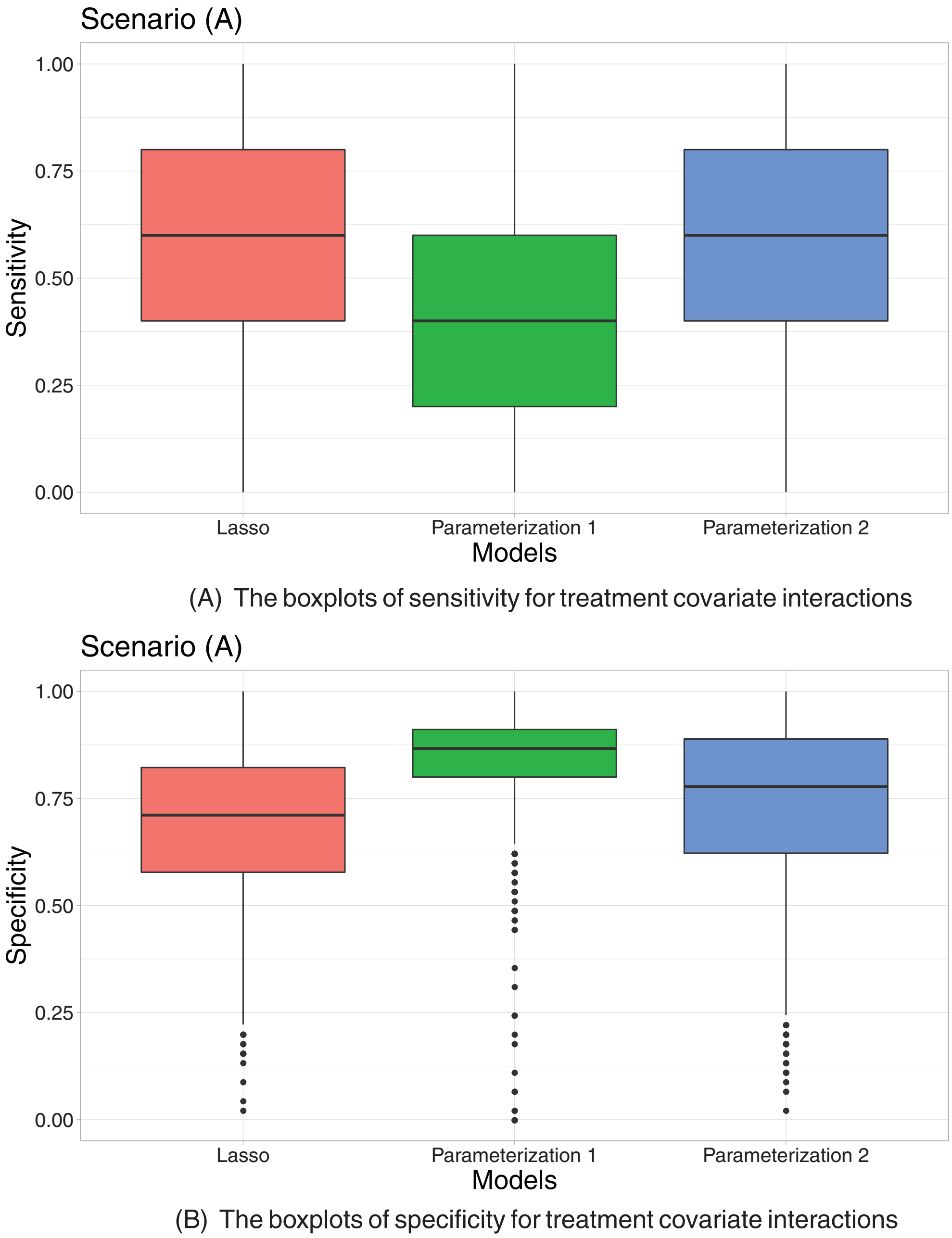
(A) The boxplots of concordance index in scenario (A)

(B) The boxplots of concordance index in scenario (B)

FIGURE 1 The comparison in risk prediction using concordance index [Colour figure can be viewed at wileyonlinelibrary.com]

References 43 and 44. The concordance index is a common metric to evaluate the risk prediction for TTE outcome, and if a pair of comparable subjects are drawn randomly from the population, it represents the probability that the subject with higher predicted risk would experience the event before the other one. For each of the methods, the penalization parameter λ should be set so that the model can be estimated. We apply 10-fold cross-validation on each 200 subjects simulated trial in search of the λ that corresponds to the highest concordance index using the out-of-sample risk predictions for these 200 subjects. Accompanying each simulated trial is an invisible trial where we generate another 10 000 subjects data under the same mechanism, serving as the validation set. We apply each determined model to this validation set to compute a concordance index as the assessment of prediction performance for each method. We repeat this process 1000 times and summarize the results in Figure 1 for scenario (A) and (B).

Figure 1A displays the performance in risk prediction via the boxplots of concordance index for each of the methods in scenario (A) where the hierarchical interaction restriction is enforced in the data generating distribution while that for scenario (B) is represented in Figure 1B. We can clearly see that in both scenarios all these methods have comparable risk



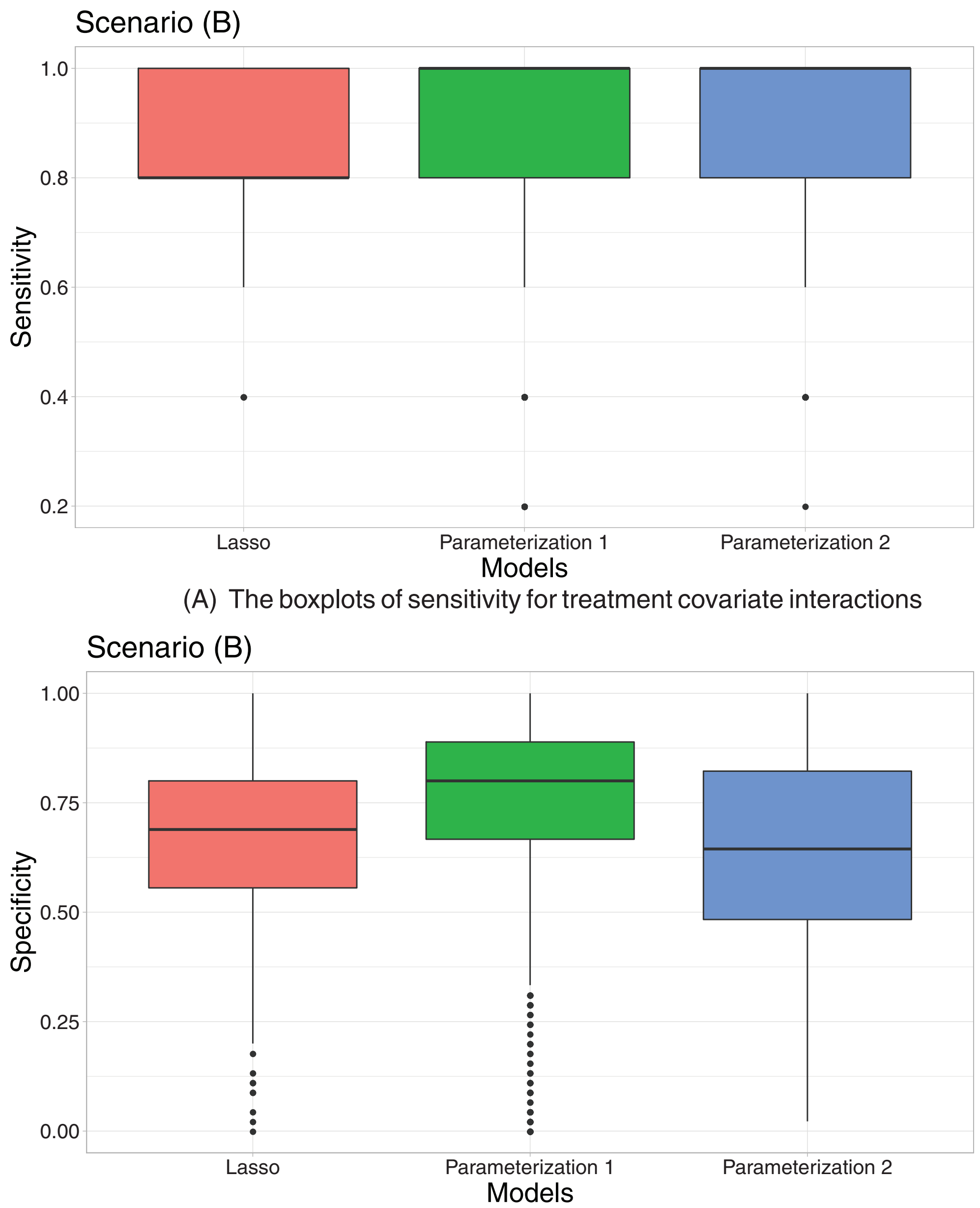
(A) The boxplots of sensitivity for treatment covariate interactions

(B) The boxplots of specificity for treatment covariate interactions

FIGURE 2 The ability to recover nonzero interactions in scenario (A) [Colour figure can be viewed at wileyonlinelibrary.com]

prediction accuracy, and on average, both parameterizations listed in Section 2.2 that incorporate interaction hierarchy have greater concordance index than Lasso. This gives us the reassurance for the proposed methods since even in scenario (B) where the truth does not support interaction hierarchy, our proposed methods can still rival Lasso in terms of risk prediction.

The main advantage of the proposed methods lies in the ability to correctly recover the significant, namely, nonzero, treatment-covariate interactions. To demonstrate this ability, we assess the sensitivity and the specificity of each model for the identification of treatment-covariate interaction terms. In addition, we also created a novel measure of global performance, called global interaction recovery cost (GIRC), to provide a single metric combining the sensitivity and specificity. Once the model is determined using cross-validation, it is straightforward to compute the sensitivity and the specificity with respect to the interaction terms. Figure 2 shows the boxplots of the sensitivity (Figure 2A) and the specificity (Figure 2B) for each method regarding the identification of treatment-covariate interactions in scenario (A). As shown in Figure 2A, the sensitivity of recovering nonzero interactions for Lasso is comparable to PS2, better than



(A) The boxplots of sensitivity for treatment covariate interactions

(B) The boxplots of specificity for treatment covariate interactions

FIGURE 3 The ability to recover nonzero interactions in scenario (B) [Colour figure can be viewed at wileyonlinelibrary.com]

PS1. However, remember that the simulation has only five nonzero treatment-covariate interactions, and the difference between the proposed methods and Lasso in terms of the average sensitivity is small, amounting to a difference of a single interaction. That means on average, Lasso and PS2 can recover three out of five nonzero interactions, while the PS1 are able to identify to out of five. When it comes to the specificity of the interaction terms as displayed in Figure 2B, the advantage of our methods compared with Lasso is clear. Both PS1 and PS2 show superior performance than Lasso. In particular, on average PS1 mark correctly around seven more treatment-covariate interaction terms as significant than Lasso. Thus, the treatment-covariate interactions identified by the proposed methods are more likely to be the true effect modifiers than those given by Lasso.

Figure 3 displays the corresponding performance of each method in the sensitivity (Figure 3A) and the specificity (Figure 3B) for scenario (B) where the truth violates the interaction hierarchy restriction. This means that both PS1 and PS2 are misspecified models. In terms of the sensitivity, the PS1 and PS2 have similar performance as Lasso, with all of them being able to recover four out of five nonzero interactions on average. While for specificity, the PS1 produce better result than PS2 and Lasso.

TABLE 2 The average global interaction recovery cost of the identification of treatment-covariate interactions for each method in scenario (A) and (B)

	PS1	PS2	Lasso
Scenario (A)	0.11	0.15	0.16
Scenario (B)	0.14	0.18	0.16

Note: PS1 and PS2 represent parameterization schemes 1 and 2, respectively. We set $C_1 = 1/2, C_2 = 1/2$.

3.3 | Global metric of performance

We propose a novel measure in order to compare the overall performance of our method in terms of both sensitivity and specificity—GIRC. This metric summarizes the sensitivity and specificity of identifying treatment-covariate interactions for the model. The sensitivity encodes the avoidance of false negative interaction terms while the specificity denotes the avoidance of false positive interaction terms. Let $\mathcal{N}_{\text{FP}}$ and $\mathcal{N}_{\text{FN}}$ denote the number of false positive and false negative treatment-covariate interaction terms, respectively. We use $\mathcal{N}$ to represent the total number of candidate interaction terms. Two types of error are associated with the false positive and false negative interaction terms, error of commission and error of omission. The error of commission is committed when the irrelevant interactions are included in the model (false positive) while the error of omission is committed when the relevant interactions are excluded from the model (false negative). Let C_1 and C_2 be the unit costs associated with committing the error of omission (ie, cost per excluded relevant interaction term) and committing the error of commission (ie, cost per included irrelevant interaction term), respectively. We assume that C_1 and C_2 are constants that may reflect monetary cost, time cost, resource cost, and so on, for the two types of error. The GIRC is constructed such that

$$\text{GIRC} = 1/\mathcal{N}\left(C_1\mathcal{N}_{\text{FN}} + C_2\mathcal{N}_{\text{FP}}\right), \tag{15}$$

where C_1 and C_2 can be normalized so that $C_1 + C_2 = 1$. Therefore, GIRC is a global metric of performance, combining the sensitivity and the specificity and informing the combined cost of recovering nonzero interaction terms. The greater the ability of capturing the significant treatment-covariate interactions is, the smaller value of GIRC should be. Assuming $C_1 = C_2 = 0.5$ where the two types of error incur the same amount of cost per term, we compute the average GIRC for each method in scenario (A) and (B), which are summarized in Table 2.

One more simulation setting is added where ζ_j in the data-generating mechanism from Scenario (A) is set to come from the distribution $U[-0.5, 0.5]$ instead of $U[-1, 1]$. This additional simulation helps us to investigate if a smaller magnitude of the interaction term will affect the performance of our proposed parameterizations vs Lasso. In this setting, PS1 is still found to exhibit the best performance with the smallest GIRC with the GIRC values being 0.09 for PS1, 0.13 for PS2, and 0.14 for Lasso, respectively. Therefore, our proposed hierarchical formulation is robust to the reduction of magnitude of the interaction terms.

In addition to varying the magnitude of the interaction terms, we also modified the sample sizes. Besides the size of 200, we set the sample size in each simulation setting to be 500. With the larger sample size, there is a reduction in GIRC for all approaches due to the gain in precision, although our proposed methods still outperform Lasso.

GIRC, as computed in this simulation study, demonstrates the superior performance of our proposed methods against Lasso, especially when there is interaction hierarchy in place, as in Scenario (A). The median ratio of GIRC comparing Lasso and PS1 is 1.29, while comparing to PS2 is 1.20. Note that we have assumed that two types of error incur equal cost, $C_1 = C_2 = 0.5$. In practice, the costs for two types of errors are likely to differ. Committing an error of commission is likely to be more costly in the sense that the resulting false positive treatment-covariate interactions may give rise to unnecessary external validation trials, and misguide the treatment recommendations. Therefore, it is often the case that $C_1 \le C_2$. As we see in Figures 2B and 3B, the advantage of using our proposed methods are seen in terms of greater specificity, that is, fewer false positive interaction terms, compared with Lasso. Table 3 shows the average GIRC when the cost of committing error of commission is twice that of committing error of omission, $C_2 = 2C_1$. Notably, under Scenario (A), the median GIRC ratio comparing Lasso and PS1 increases to 1.75. We also consider the rare case when the cost of

TABLE 3 The average global interaction recovery cost of the identification of treatment-covariate interactions for each method in scenario (A) and (B)

	PS1	PS2	Lasso
Scenario (A)	0.13	0.18	0.20
Scenario (B)	0.18	0.24	0.20

Note: PS1, PS2 represents parameterization scheme 1 (6), 2 (7), respectively. $C_1 = 1/3, C_2 = 2/3$.

TABLE 4 The average global interaction recovery cost of the identification of treatment-covariate interactions for each method in scenario (A) and (B)

	PS1	PS2	Lasso
Scenario (A)	0.09	0.11	0.12
Scenario (B)	0.10	0.12	0.11

Note: PS1, PS2 represents parameterization scheme 1 (6), 2 (7), respectively. $C_1 = 2/3, C_2 = 1/3$.

TABLE 5 The averaged computing time (in seconds) to converge for each iteration in each method in Scenario (A) and (B)

	PS1	PS2	Lasso
Scenario (A)	24.6	24.9	32.8
Scenario (B)	25.2	16.6	28.8

Note: PS1, PS2 represents parameterization Scheme 1 (6), Parameterization Scheme 2 (7), respectively.

false negative is higher than that of false positives where we set $C_1 = 2C_2$. The summaries are provided in Table 4. We notice that PS1 still dominates in both scenarios despite the difference between different models becoming smaller.

For the simulation studies, we compare the time of convergence for PS1, PS2, and Lasso, with results being summarized in Table 5. We apply the SPG algorithm for all the methods in order to have a fair comparison. We find that under these two scenarios, the PS1 and PS2 tend to have faster convergence than Lasso.

As a sensitivity analysis, we also varied the censoring rate in the data generation process. The considered censoring rates are 0.1, 0.3, and 0.5. The results are shown in Table 6. As the censoring rate increased, the GIRC increased. However, PS1 still has the better performance compared with both PS2 and Lasso, in both scenarios (A) and (B).

4 | DATA APPLICATION

4.1 | Data description

We evaluated our proposed methods on a completed, placebo-controlled, randomized trial (SOLVD-T) that tested the efficacy of an experimental drug–enalapril, the angiotensin-converting-enzyme inhibitor, for treating chronic heart failure patients.[29] 1284 patients were assigned randomly to the control arm while 1285 to the treatment arm. The primary outcome of the trial is the time to hospitalization or death. We assumed noninformative censoring, and around 47% of the outcome were censored. There are 23 candidate effect modifiers, including baseline age, gender, New York Heart Association function status, sodium level, creatinine level, and so on. The goal of our proposed methods for this study is to select

TABLE 6 The average global interaction recovery cost of the identification of treatment-covariate interactions for each method in scenario (A) and (B) for censoring rates of 0.1, 0.3, and 0.5

Censoring rate: 0.1			
	PS1	**PS2**	**Lasso**
Scenario (A)	0.11	0.15	0.16
Scenario (B)	0.14	0.18	0.16
Censoring rate: 0.3			
	PS1	**PS2**	**Lasso**
Scenario (A)	0.12	0.17	0.19
Scenario (B)	0.12	0.17	0.17
Censoring rate: 0.5			
	PS1	**PS2**	**Lasso**
Scenario (A)	0.15	0.20	0.24
Scenario (B)	0.15	0.20	0.22

Note: PS1, PS2 represents parameterization schemes 1 (6) and 2 (7), respectively. $C_1 = 1/2, C_2 = 1/2$.

a subset of these candidates that are predicted to have nonzero interactions with treatment, and to provide the estimates. Missing values in the dataset were imputed, where missing at random was assumed. For example, if Z_1 has missing values, we regress Z_1 on the observed values of all other covariates $Z_2, \ldots, Z_p$ so that the missing values are filled by the predictive value from the regression. This imputation is carried out one variable at a time.

4.2 | Direct application

We applied our proposed methods as well as unstructured Lasso to the SOLVD-T trial, and provided a heatmap in Figure 4 showing the identified treatment-covariate interactions. Each column represents a method while each row shows a candidate effect modifier. A complete list of description of these variables is given in the Supplementary Material. The black area indicates which candidates are identified by each model, whereas the blank area shows no treatment-covariate interaction. It can be observed that 5, 7, and 11 variables were predicted to influence the patient's response to treatment, respectively, by PS1 (6), PS2 (7) and Lasso. The baseline covariates, lvef, beat, and creatinine were found by all three methods to significantly modify the effect of enalapril on the survival outcome. LVEF was by far the strongest treatment effect modifier, with greater beneficial effect seen at the lower values of ejection fraction. Similar findings were also reported in the study of heterogeneous treatment response using the same trial data.[45] We are unable to assess the sensitivity/specificity of the methods in this real trial since the true effect modifiers remain unknown, but it serves a good purpose to illustrate the use of our methodology for a real data application.

4.3 | Friedman's randomly generated functions

Additional simulations were conducted based on the SOLVD trial data to further evaluate the performance of our proposed methods compared with Lasso, using randomly generated nonlinear regression functions. Here, we introduced two layers of model misspecification while keeping the original relationship among covariates unchanged. First, we considered an accelerated failure time (AFT) model where the TTE outcome is directly linked to the covariates and the treatment assignment through

$$\log T = m(A, Z) + W. \tag{16}$$

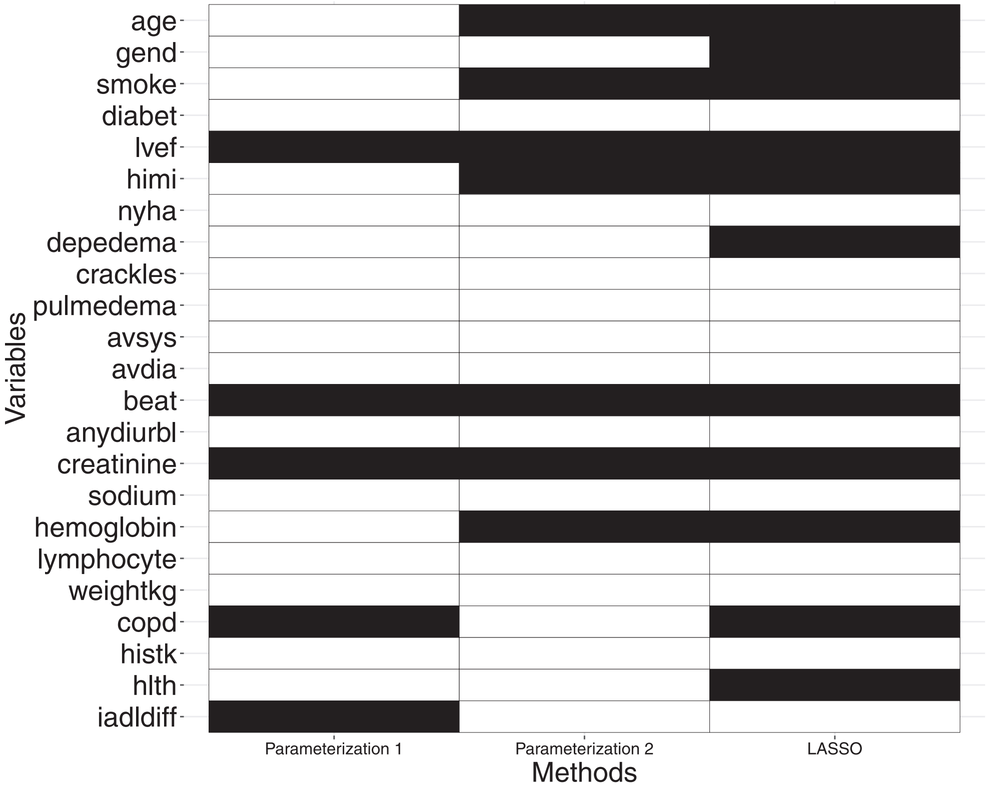


FIGURE 4 Treatment-covariate interaction recovery map for the proposed methods and Lasso

W is a residual term. The AFT model is considered a useful alternative to the Cox model in survival analysis.[46-48] Second, we tested the performance of our methods using complex nonlinear functions that involve higher order interactions. To generate these nonlinear functions, we adapted the approach developed by Reference 49 originally to evaluate the performance of gradient boosted regression trees. The similar adaptation was also applied in Reference 45.

The random regression functions $m(A, Z)$ were generated through

$$m(A_i, Z_i) = F_0(Z_i) + A_i * F_1(Z_i), \tag{17}$$

where F_0 represents the main effect function and F_1 the effect modification function. These are defined as

$$F_0(Z_i) = \sum_{k=1}^{10} a_{1k} h_{1k}(v_{1k}) \tag{18}$$

$$F_1(Z_i) = \sum_{k=1}^{10} a_{2k} h_{2k}(v_{2k}). \tag{19}$$

The coefficients a_{1k} and a_{2k} are assumed to follow Uniform(−1, 1) and Uniform(0, 0.5), respectively. For main effect function F_0 and for each k, we randomly select a subset of Z_i, v_{1k}, whose sizes are determined by $\min(r_k + 1, 3)$, where $r_k \sim$ Exponential(0.5). For function F_1, the subset sizes are fixed at 3. The function $h_{jk}(v_{jk})$ is defined such that

$$h_{jk}(v_{jk}) = \exp\left\{-\frac{1}{2}(v_{jk} - \mu_{jk})'\Sigma_{jk}(v_{jk} - \mu_{jk})\right\}. \tag{20}$$

The vector μ_{jk} is generated such that each of its element follows standard normal distribution, and the random matrix Σ_{jk} is constructed as $\Sigma_{jk} = U_{jk}\Lambda_{jk}U'_{jk}$ where U_{jk} is a random orthogonal matrix and Λ_{jk} is a diagonal matrix with the square root of each element coming from Uniform(0.1, 2). In order to generate the random orthogonal matrix U_{jk}, we first generated a random Gaussian matrix where every element follows a standard normal distribution and applied the

TABLE 7 The average global interaction recovery cost of the identification of treatment-covariate interactions for each method in Friedman simulation

	PS1	PS2	Lasso
Friedman simulation	0.06	0.10	0.10

Note: PS1 and PS2 represent parameterization scheme 1 and 2, respectively. $C_1 = 1/2, C_2 = 1/2$.

TABLE 8 The mean partial specificity of the identification for treatment-covariate interactions for each method based on the SOLVD-T trial

	PS1	PS2	Lasso
Mean	99.5%	79.5%	62.6%

Note: PS1 and PS2 represent parameterization scheme 1 and 2, respectively.

eigen-decomposition to obtain the eigenvectors matrix as U_{jk}. The three variables representing ejection fraction, heart beat and the level of creatinine from the dataset were chosen to be the prognostic and predictive factors. The TTE outcome was thus generated employing AFT model with the above random nonlinear regression functions, using independent censoring which follows a Uniform distribution.

Table 7 summarizes the average GIRC for each method over 1000 simulations. PS1 (6) stood out as having the lowest averaged GIRC among all the methods. On average, this method can identify two out of three prespecified terms as predictive variables and provide highly desirable specificities equal to 96%. Remember that two layers of model misspecification were present in these simulations while the original structure and relationship among covariates remained unchanged. Therefore, with these conditions in mind, the results clearly vouch for the robust performance of PS1 (6) under varying conditions. By contrast, there is a drop in the performance of PS2 (7), although it is still comparable to Lasso. This indicates that PS2 might be more sensitive to the underlying modeling assumptions (eg, proportional hazard assumption).

4.4 | More real data based simulations

Since the sensitivity / specificity cannot be evaluated on the real data because the truth is seldom known, we engineered a real data-based design to assess what we have termed as the “partial specificity” of these methods. We created m noise variables that have no association at all with the response and added to the original data. We then fit each method to this expanded dataset. This “partial specificity” was thus defined to be the percentage of the noise variables that are correctly identified by the method as having no interaction with the treatment. We set $m = 25$ and repeat the procedure 100 times. A summary of the averaged “partial specificity” for each method is provided in Table 8. An improvement regarding this partial specificity can be seen for both of our proposed methods over Lasso. Lasso had the lowest value 62.6%, which says that for every 100 noise variables added, the Lasso would falsely identify 37 as effect modifiers. Surprisingly, PS1 (6) nearly rejected all the noise variables with close to 100% partial specificity, whereas PS2 (7) had a partial specificity of around 80%. This observation, aligned with the real data based simulation study in Section 4.3, shows that when the true data generating distribution is unknown, as in this real trial, PS1 (6) seems more robust to the modeling assumptions than PS2 (7).

5 | DISCUSSION

Understanding how patients respond differently to treatment is key to translating clinical trial findings. Prediction models for outcome prognosis are common in clinical research, however, models that predict treatment response heterogeneity in terms of individual baseline characteristics are much less common. Such models can be useful for answering the question of which characteristics are predictive of the treatment response. We provided a general prediction method to assess treatment response heterogeneity by extending Lasso to honor the interaction hierarchy restriction. We gave a clear interpretation on how individual characteristics affect treatment response by modeling pairwise interactions between treatment and covariates. We further extended the work of Reference 22 by relaxing the constraint that the treatment-covariate interactions are proportional to the main effects. Our proposed methods are able to automatically screen a larger number of candidate effect modifiers with the aid of parameter regularization inherited from Lasso in an effort to capture those with nonzero treatment-covariate interactions that strike a good balance between false positives and false negatives. The simulation study in Section 3 and the real data examples shown in Section 4 demonstrated superior performance of our proposed methods against Lasso with regards to the ability to correctly identify the nonzero interactions.

In terms of choosing a parameterization scheme in practice, we recommend to use PS1 (6) as the default choice because it exhibited the best performance under a variety of settings. When the parametric model assumption and interaction hierarchy restriction can well approximate the true data generating distribution, PS2 (7) demonstrated a higher sensitivity to capture nonzero interactions than PS1 and a higher specificity over Lasso as shown in Section 3.

Despite the focus of this article on TTE outcome, our methods can easily be applied to other types of endpoints, for example, continuous or binary. Only the objective function (the log-likelihood) needs to be rewritten. For binary outcome this would be a binomial likelihood, while for continuous outcome this would be a Gaussian likelihood. All the constraints would remain the same. The same optimization algorithm SPG can be used.

There are some limitations with applying our proposed methods for the analysis of treatment response heterogeneity. These limitations in turn open the door for future extensions and improvements. First of all, the article focuses on the prediction side rather than inference (eg, confidence intervals and *P*-values) because it is difficult to draw inferences on Lasso type of problem and very limited approaches are available for this purpose. Inference for our models is even more difficult than the standard Lasso formulation because of the presence of constraints. Second, our proposed methods are built upon (semi)parametric model assumptions, such as those in the Cox model (1). Reference 50 presents a stagewise estimation with generalized estimating equations to select interactions for the clustered data. The merit of using generalized estimating equations are twofold: (1) it does not require a parametric model assumption, and (2) it allows for the analysis of longitudinal/clustered data. It is possible to adapt our proposed methods to their stagewise procedure in order to relax the parametric model assumption. This approach is also likely to be more computationally efficient than the SPG method. Various methods exist to solve the constrained optimization problem. One area of future work would be to find algorithms with better computational efficiency.

Finally, the proposed methods, as they stand, are not quite ideal for making personalized treatment recommendations. In fact, our main goal is to develop a method that can identify predictors of treatment response in a way that delivers a good balance between false positives and false negatives. If the goal is personalized treatment choice, then the emerging area of machine learning methods such as causal random forests and Bayesian additive regression trees appear to be better suited. Nonetheless, our proposed methods have a role to play in the exploratory, secondary analyses of large, phase 3 clinical trials. They facilitate the assessment of heterogeneity in patient's response to the experimental treatment. The identified treatment effect modifiers should be externally and independently validated in subsequent studies.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were collected in this study.

REFERENCES

1. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995;345(8965):1616-1619.
2. Bailey KR. Generalizing the results of randomized clinical trials. *Control Clin Trials*. 1994;15(1):15-23.
3. Kent D, Hayward R. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA*. 2007;298(10):1209-1212.
4. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst*. 2007;99(13):1036-1043.
5. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. *Clin Trials*. 2008;5(3):181-193.
6. Barker A, Sigman C, Kelloff G, Hylton N, Berry D, Esserman L. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther*. 2009;86(1):97-100.
7. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clin Cancer Res*. 2010;16(2):691-698.
8. Lee JJ, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. *Clin Trials*. 2010;7(5):584-596.

9. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov*. 2011;1(1):44-53.
10. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*. 2011;12(2):270-282.
11. Lai TL, Lavori PW, Liao OYW. Adaptive choice of patient subgroup for comparing two treatments. *Contemp Clin Trials*. 2014;39(2):191-200.
12. Xu Y, Trippa L, Müller P, Ji Y. Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Stat Biosci*. 2016;8(1):159-180.
13. Ohwada S, Morita S. Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharm Stat*. 2016;15(5):420-429.
14. Spencer AV, Harbron C, Mander A, Wason J, Peers I. An adaptive design for updating the threshold value of a continuous biomarker. *Stat Med*. 2016;35(27):4909-4923.
15. Henderson NC, Louis TA, Wang C, Varadhan R. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Serv Outcome Res Methodol*. 2016 Sep;16(4):213-233.
16. Abrevaya J, Hsu YC, Lieli RP. Estimating conditional average treatment effects. *J Bus Econ Stat*. 2015;33(4):485-505.
17. Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. *Proc Natl Acad Sci*. 2016;113(45):12673-12678.
18. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. Paper presented at: Proceedings of the 34th International Conference on Machine Learning-Vol. 70, 2017:3076-3085; JMLR. org.
19. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.
20. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.
21. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics*. 1991;47(3):871-881.
22. Kovalchik SA, Varadhan R, Weiss CO. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Stat Med*. 2013;32(28):4906-4923.
23. Follmann DA, Proschan MA. A multivariate test of interaction for use in clinical trials. *Biometrics*. 1999;55(4):1151-1155.
24. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Ann Statist*. 2013;41(3):1111-1141.
25. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
26. Cox DR. Interaction. *Int Stat Rev / Revue Internationale de Statistique*. 1984;52(1):1-24.
27. Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. *Stat Methods Med Res*. 2013;22(5):493-504.
28. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc*. 2014;109(508):1517-1532.
29. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med*. 1991;325(5):293-302.
30. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B (Methodol)*. 1972;34(2):187-220.
31. Chipman H. Bayesian variable selection with related predictors. *Can J Stat / La Revue Canadienne de Statistique*. 1996;24(1):17-36.
32. Nelder JA. A reformulation of linear models. *J Royal Stat Soc Ser A (General)*. 1977;140(1):48-77.
33. Peixoto JL. Hierarchical variable selection in polynomial regression models. *Am Stat*. 1987;41(4):311-313.
34. Birgin EG, Martínez JM, Raydan M. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J Optim*. 2000;10(4):1196-1211.
35. Raydan M. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J Optim*. 1997;7(1):26-33.
36. Bertsekas D. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans Autom Control*. 1976 Apr;21(2):174-184.
37. Birgin E, Martínez JM, Raydan M. Spectral projected gradient methods: review and perspectives. *J Stat Softw*. 2014;60(3):1-21.
38. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016.
39. Varadhan R, Gilbert P. BB: an R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *J Stat Softw*. 2009;32(4):1-26.
40. Barzilai J, Borwein JM. Two-point step size gradient methods. *IMA J Numer Anal*. 1988;8(1):141-148.
41. Varadhan R, Roland C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand J Stat*. 2008;35(2):335-353.
42. Kraft D. A software package for sequential quadratic programming. technical report DFVLR-FB 88-28; 1988.
43. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
44. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004;23(13):2109-2123.
45. Henderson NC, Louis TA, Rosner GL, Varadhan R. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. June 2017. ArXiv e-prints.
46. Louis TA. Nonparametric analysis of an accelerated failure time model. *Biometrika*. 1981;68(2):381-390.
47. Robins J, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time- dependent covariates. *Biometrika*. 1992;79(2):311-319.
48. Wei LJ. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat Med*. 1992;11(14-15):1871-1879.

49. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist*. 2001;29(5):1189-1232.
50. Vaughan G, Aseltine R, Chen K, Yan J. Stagewise generalized estimating equations with grouped variables. *Biometrics*. 2017;73(4):1332-1342.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Du Y, Chen H, Varadhan R. Lasso estimation of hierarchical interactions for analyzing heterogeneity of treatment effect. *Statistics in Medicine*. 2021;1–17. https://doi.org/10.1002/sim.9132