Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

A Generalized Quantile Tree Method for Subgroup Identification

Xiang Peng & Huixia Judy Wang

To cite this article: Xiang Peng & Huixia Judy Wang (2022) A Generalized Quantile Tree Method for Subgroup Identification, Journal of Computational and Graphical Statistics, 31:3, 824-834, DOI: 10.1080/10618600.2022.2032723

To link to this article: https://doi.org/10.1080/10618600.2022.2032723

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 31, NO. 3, 824–834
https://doi.org/10.1080/10618600.2022.2032723

Taylor & Francis
Taylor & Francis Group

A Generalized Quantile Tree Method for Subgroup Identification

Xiang Peng and Huixia Judy Wang

Department of Statistics, George Washington University, Washington, DC

ABSTRACT

One primary goal of subgroup analysis is to identify subgroups of subjects with differential treatment effects. Existing methods have focused on the mean treatment effect and may be ineffective when the two distributions differ in scales or in the upper or lower tails. We develop a new generalized quantile tree method for subgroup identification. The method first uses quantile rank score tests to select split variables and then estimates the split point by minimizing a composite quantile loss. The proposed split rule is free of variable selection bias and robust against outliers and heavy-tailed distributions. In addition, we introduce a generalized quantile treatment effect estimator and a testing method for the selection and confirmation of predictive subgroups. Simulation shows that the proposed method gives more accurate subgroup identification than existing methods for cases with heteroscedastic or heavy-tailed errors. The practical value of the method is demonstrated through the analysis of an AIDS clinical trial data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Revised January 2022

KEYWORDS
Quantile treatment effect; Rank score; Recursive partitioning; Regression tree; Subgroup analysis.

1. Introduction

For many diseases, treatment efficacy may vary across individuals. Rather than searching for a beneficial treatment for all subjects, it would be more practical to identify subgroups with differential treatment effects. With the assessment of heterogeneous effects, subgroup analysis explores how patients with particular characteristics respond to a given treatment, and therefore, provides guidance for tailored therapies. Furthermore, for failed clinical trials, subgroup identification has attracted increasing attentions as it helps examine hidden beneficial effects among subpopulations. For example, Lefitolimod, a drug developed by Mologen AG for the treatment of HIV infections and various cancers, has been shown to prolong survival to a subgroup of patients with extensive-stage small-cell lung cancer (Mologen 2018).

Various methods have been proposed to facilitate the procedure of subgroup identification. The model-based clustering approaches of Imai and Ratkovic (2013), Cai et al. (2011) and Zhao et al. (2013) require fitting a model with prespecified main and interaction terms. The resulting model is then used to evaluate subject-specific treatment differences, and the target subgroup is formed by including individuals with estimated treatment effects exceeding a certain threshold. For such approaches, it is often difficult to interpret the features of the selected group. Alternative methods are developed under the formulation of change set regression; see, for example, Chen et al. (2015) and Huang et al. (2017). These methods identify subgroups with improved treatment effects by examining the significance of the treatment-covariate interactions, but they do not account for the potential prognostic effects of the predictive variables. Bayesian methods have also been employed to evaluate predictive effects in complex regression models; see Jones et al. (2011), Sivaganesan, Laud, and Müller (2011), Gu, Yin, and Lee (2013) and Berger, Wang, and Shen (2014). These methods can easily incorporate prior information but they are often computationally expensive.

Tree-based methods have drawn more attention in recent years for subgroup analysis due to the appealing feature that treatment-covariate interactions can be identified without the need to prespecify the interaction terms in the model. Rather than fitting a global model over the entire data space, a regression tree applies a split rule recursively to partition the space into small regions. Therefore, the interactions can be identified in a more convenient way by fitting local models. This is appealing especially when the study includes a large number of features that may interact with the treatment in complicated and nonlinear ways. Furthermore, the resulting tree structure can help with data visualization and interpretation, since subgroups are defined naturally by the path from root to leaf. Most existing tree-based methods (Breiman et al. 1984; Su et al. 2009; Foster, Taylor, and Ruberg 2011; Lipkovich et al. 2011; Loh, He, and Man 2015; Seibold, Zeileis, and Hothorn 2016) are confined to searching for predictive subgroups via the assessment of mean or median treatment effects. However, several clinical studies have shown that the treatment and control distributions may differ not in the center, but in scales or at the lower or upper tails; see, for example, Keystone et al. (2004) and Kremer et al. (2006). For such data, the mean or median methods may be ineffective to identify subgroups with differential treatment effects.

To capture different forms of heterogeneous treatment effects, we develop a new generalized quantile tree method

CONTACT Huixia Judy Wang *judywang@gwu.edu* Department of Statistics, George Washington University, Washington, DC.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

for subgroup identification. The method first uses quantile rank score tests to determine the split variable, and this approach is free of variable selection bias and able to capture noncentral difference induced by predictive variables. After this, the method estimates the associated split points for selected variables by minimizing a composite quantile loss. The proposed split rule not only reduces computational efforts but is also robust against outliers and heavy-tailed distributions. Furthermore, we introduce a generalized quantile treatment effect (GQTE) estimator to measure and test the treatment effect at each node. Compared with conventional testing procedures, the GQTE test is adaptive to detecting treatment effects of various forms.

The remainder of the article is organized as follows. Section 2 introduces the proposed method, including the tree-building and the evaluation of treatment effects. We assess the performance of the proposed method through simulation studies in Section 3, and apply the method to an AIDS clinical trial dataset in Section 4. Section 5 concludes the article with some discussion. Technical proofs and some additional simulation results are provided in the supplementary materials.

2. Proposed Method

2.1. Notation and Motivation

Let $\{y_i, z_i, \mathbf{x}_i\}$ be the observed data of subject i, where y_i denotes the continuous response, z_i is the binary treatment indicator and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the p-dimensional covariate vector. The covariates can be either categorical or continuous. An important part of subgroup analysis is to identify covariates that define patient subgroups with more (or less) beneficial treatment effects, which are referred to as predictive variables in the medical literature (Mehta et al. 2010; Italiano 2011). In contrast, a variable is prognostic if it provides information on the outcome distribution and does not interact with the treatment. Under the regression setup, prognostic variables have marginal effects on the response for untreated subjects, while predictive covariates are involved in treatment-covariate interaction terms, which can be used to define subgroups.

For example, consider the following model

$$Y = 1.56 + 0.8Z + 0.75I(X_1 > 0.45) + 1.6I(X_2 < 1.34)Z + \epsilon,$$

where $\epsilon \sim N(0, 1)$. Then X_1 is a prognostic variable since it has marginal effects on the response distribution for untreated patients. On the other hand, X_2 is a predictive variable that interacts the treatment indicator Z. The predictive variable X_2, together with the cutoff 1.34, defines a subgroup with enhanced treatment effect, that is, $X_2 < 1.34$. In this project, we focus on experimental studies with balanced designs. Our goal of subgroup identification is to identify subgroups with differential treatment effects that are characterized by predictive variables and the associated cutoff values.

In general, there are three important steps when applying tree-based approaches for subgroup identification. The first two steps are split variable selection and split point estimation, which are employed repeatedly to dichotomize the data at each node. Once the tree is generated by applying the split rule recursively, the next step is to select and confirm the predictive subgroups by evaluating the treatment effect at each node.

Most existing tree-based methods search for subgroups by assessing the mean or median treatment effects, which may overlook some important differences in the scales or tails of the distributions. Furthermore, tests based on center measurements may suffer from low power for cases with outlying observations or data from skewed distributions. For example, as demonstrated in a rheumatoid arthritis clinical trial study (van der Heijde et al. 2006; He et al. 2010), the changes in Total Sharp Scores, the primary measurements of the treatment effects, are nearly identical for about 75% of patients and differ from patients with the most progressive diseases on the right tail; see Figure 1. For such data, existing tree-based methods may fail to detect the tail difference.

To capture different forms of treatment effects, we develop a new regression tree method for subgroup identification based on quantile rank scores and generalized quantile treatment effect. The method uses quantile-based approaches for all three steps: split variable selection, split point estimation, and the selection and confirmation of predictive subgroups. Compared with existing approaches, the method can better adapt to different forms of treatment effects, and is robust against outliers and heavy-tailed distributions. Hereafter we refer to the proposed procedure as the generalized quantile (GQ) method. Below we will introduce the proposed method for the three steps separately.

2.2. Split Variable Selection

For tree-based methods, the tree structure is determined by the selected split variables and the corresponding split points. Existing methods are either based on exhaustive searches (Su et al. 2009; Lipkovich et al. 2011), which are computationally intensive and often select variables that allow more splits, or focus on mean measurements (Loh 2002; Zeileis, Hothorn, and Hornik 2008; Loh, He, and Man 2015; Seibold, Zeileis, and Hothorn 2016), which are susceptible to outliers. To overcome such limitations, we propose a quantile-based splitting procedure, where a quantile rank score test is used for choosing the split variable and a composite quantile method is then used for estimating the split point. The method is computationally convenient, free of selection bias and robust against outliers.

We propose to detect treatment-covariate interactions induced by predictive covariates and select the splitting variables by adapting the rank score test in Gutenbrunner et al. (1993) and Koenker (2010). Suppose that the candidate covariate X is a categorical variable with L levels. A continuous covariate can be discretized, for example, by the sample quartiles and converted to a category variable with $L = 4$. We consider the following quantile regression model:

$$Q_\tau(y_i|z_i, x_i) = \alpha(\tau) + z_i\beta(\tau) + \boldsymbol{l}_i^T\boldsymbol{\eta}(\tau) + z_i\boldsymbol{l}_i^T\boldsymbol{\delta}(\tau), \quad (1)$$

where $Q_\tau(y_i|z_i, x_i)$ is the τth conditional quantile of y_i given the covariates with $0 < \tau < 1$ being the quantile level, $\boldsymbol{l}_i = (I(x_i = 2), \ldots, I(x_i = L))^T \in \mathbb{R}^{L-1}$ is an indicator vector representing the level of X for subject i, and $(\alpha(\tau), \beta(\tau), \boldsymbol{\eta}(\tau), \boldsymbol{\delta}(\tau))$ are unknown coefficients. A popular class that leads to model (1)

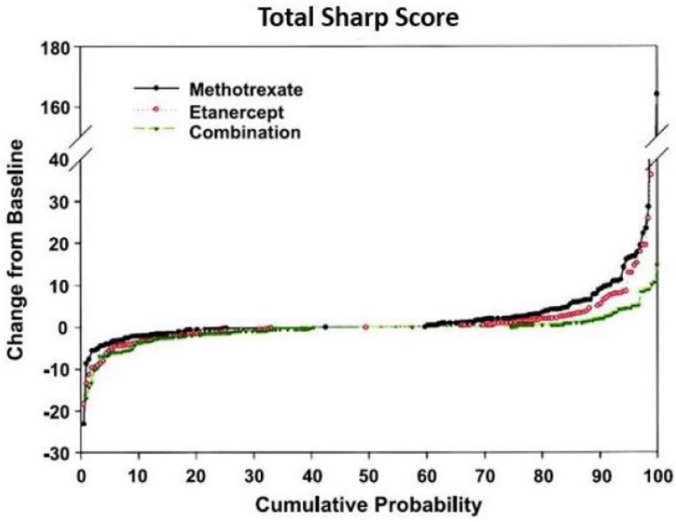

Figure 1. The changes from baseline in Total Sharp Scores of the three treatment groups.

is the location-scale-shift model

$$y_i = \alpha + z_i\beta + \boldsymbol{l}_i^T\boldsymbol{\eta} + z_i\boldsymbol{l}_i^T\boldsymbol{\delta} + \sigma_i\epsilon_i, \tag{2}$$

where ϵ_i are iid from the distribution function F, and $\sigma_i = (1, z_i, \boldsymbol{l}_i^T, z_i\boldsymbol{l}_i^T)\boldsymbol{\gamma}$ is the positive scale function that is left unspecified.

Hereafter we focus on this class to explain the properties of the rank score test. To detect the interaction effect between X and the treatment, we can test the following hypotheses H_0 : $\boldsymbol{\delta}(\tau) = \mathbf{0}, \tau \in \mathcal{T}$ for some index set $\mathcal{T} \subset (0,1)$ against H_a : $\boldsymbol{\delta}(\tau) \neq \mathbf{0}$.

Under the null model, regression rank scores are defined as

$$\hat{\mathbf{a}}(\tau) = \text{argmax}\{\mathbf{a}^T\mathbf{y} | D_0^T\mathbf{a} = (1-\tau)D_0^T\mathbf{1}, \mathbf{a} \in [0,1]^n\},$$

where $D_0 = \{1, z_i, \boldsymbol{l}_i^T\}_{i=1}^n$ is the design matrix under H_0. Let $\hat{\epsilon}_i(\tau)$ be the residuals obtained under H_0. Then $\hat{a}_i(\tau) = I\{\hat{\epsilon}_i(\tau) > 0\}$ for $\hat{\epsilon}_i(\tau) \neq 0$ and $\in (0,1)$ for $\hat{\epsilon}_i(\tau) = 0$. The vector $\hat{\mathbf{a}}(\tau)$ is the dual solution to the quantile regression problem for fitting Model (1) under H_0, and can be interpreted as "signed residuals." Define

$$\hat{b}_i^{\phi} = -\int_0^1 \phi(u)d\hat{a}_i(u), \tag{3}$$

where $\phi(\cdot)$ is some score function on $(0,1)$. When $\phi(u) = u - 1/2$, $\hat{b}_i^{\phi}$ yields "ranks" of the observations. One can choose different score functions to get more general notions of "ranks." For example, the quantile sign score $\phi(u) = \text{sgn}(u-\tau)/2 + (\tau - 1/2)$, yielding $\hat{b}_i^{\phi} = \tau - I\{\hat{\epsilon}_i(\tau) < 0\}$.

To test for the interaction effect, we consider the test statistic

$$T_n = S_n^T Q_n^{-1} S_n / A^2(\phi),$$

where $S_n = (D_1 - \hat{D}_1)^T\hat{b}^{\phi}, Q_n = (D_1 - \hat{D}_1)^T(D_1 - \hat{D}_1), \hat{D}_1 = D_0(D_0^T\Gamma^{-1}D_0)^{-1}D_0^T\Gamma^{-1}D_1, \Gamma = \text{diag}(\sigma_i)$ and $D_1 = \{\boldsymbol{l}_i^T z_i\}_{i=1}^n$, $A(\phi) = \int_0^1\{\phi(u) - \bar{\phi}\}^2du$ and $\bar{\phi} = \int_0^1 \phi(u)du$. Koenker and Machado (1999) showed that T_n is asymptotically χ^2_{L-1} under H_0, and it converges to a noncentral χ^2 distribution under the local alternative $H_n : \boldsymbol{\delta}_n(\tau) = \boldsymbol{\delta}_0(\tau)/\sqrt{n}$, where $\boldsymbol{\delta}_0(\tau)$ is a fixed continuous function for $\tau \in [0,1]$.

At each node, we apply the rank score test with a chosen score function to each candidate covariate, and then choose the split variable as the one that yields the most significant interaction effect, that is, gives the smallest p-value. The testing procedure has three main advantages. First, it is computationally convenient since it requires estimation only under the null hypothesis. Second, the rank score test has shown robustness to the error heteroscedasticity in various setups (Kocherginsky, He, and Mu 2005; Wang and Fygenson 2009), therefore, we can implement the testing procedure under the working iid error model (Model (2) with $\sigma_i \equiv 1$), which further reduces the computational effort. Third, by examining the noncentrality parameters of the rank score test under local alternatives, we can choose the score function to optimize the power and thus, accommodate different types of signals in an efficient way. The following proposition summarizes the optimal scores for the location-scale-shift models.

Proposition 1. Suppose that model (2) holds and the distribution function F has a strictly positive and continuous density f. Then

(i) for location-shift local alternatives with $\boldsymbol{\delta}_0(\tau) = \boldsymbol{\delta}_0$, the optimal score function is

$$\phi(u) \propto -\frac{f'}{f}\{F^{-1}(u)\}. \tag{4}$$

(ii) for scale-shift local alternatives with $\boldsymbol{\delta}_0(\tau) = \boldsymbol{\delta}_0 F^{-1}(\tau)$, the optimal score function is

$$\phi(u) \propto -\left[1 + F^{-1}(u)\frac{f'}{f}\{F^{-1}(u)\}\right]. \tag{5}$$

Remark 1. Proposition 1 suggests that the optimal score function depends on both the model and the error distribution, and any constant rescaling does not change the efficiency of score functions. Below we summarize the optimal score function for several scenarios.

(i) Location-shift model with Normal error: $\phi_N(u) = \Phi^{-1}(u)$, referred to as the Normal score.
(ii) Location-shift model with Logistic error: $\phi_W(u) = u - 1/2$, referred to as the Wilcoxon score.
(iii) Scale-shift model with Normal error: $\phi_{NS}(u) = \{\Phi^{-1}(u)\}^2 - 1$, referred to as the Normal scale score.
(iv) Scale-shift model with W error: $\phi_W(u) = u - 1/2$, where W has cdf $F_W(x) = (1+cx^{-1/2})^{-1} I(x > 0)$ for some $c > 0$.

For cases with signals on the tails, for example, $\beta(\tau)$ is nonzero only for $\tau > \tau_0$ for some fixed quantile level $\tau_0 \in (0, 1)$, there is no closed expression for the optimal score. However, our numerical studies and the simulation in Koenker (2010) suggest that trimmed scores, such as the half Normal scale score $\phi_{HNS}(u) = \phi_{NS}(u) I(0.5 < u < 1)$ and the trimmed Wilcoxon score $\phi_{TW}(u) = \phi_W(u) I(0.6 < u < 0.95)$, are often more efficient than nontrimmed scores. For real data, it could be challenging to identify the model type and error distribution. Therefore, we suggest an adaptive approach, which chooses the splitting variable based on the smallest p-value from multiple score functions with Bonferroni adjustment. In our numerical implementation, we consider the following four scores: Normal, Wilcoxon, half Normal scale, and trimmed Wilcoxon. Our studies show that the adaptive approach can accommodate different types of models and distributions, and it performs competitively well with the method based on the optimal scores; see details in Section 3.

2.3. Split Point Estimation

The second step of the data-splitting procedure is to estimate the cutoff point (or set) of the selected split variable, which can be viewed as a threshold regression problem. Most existing methods estimate the cutoff value by minimizing the sum of squared errors (SSE) obtained from mean regression models fitted to two child nodes, and thus, may be negatively affected by outliers or heterogeneity. To overcome such limitations, we consider estimating the split point by minimizing a composite quantile loss. This method combines information from multiple quantiles and thus, often leads to more stable and efficient estimation of the split point; see some related discussion under a different context in Zou and Yuan (2008). Let $\tau_1 < \cdots < \tau_K$ be a given grid of quantile levels. For simplicity, we consider the case where the splitting variable X is continuous; the method can also be used to dichotomize discrete splitting variables. We want to estimate the split point t to partition the data into the left node: $\{X \le t\}$ and the right node: $\{X > t\}$. The proposed composite quantile estimator of the split point is defined as

$$\hat{t} = \underset{c}{\operatorname{argmin}} \sum_{x_i \le c} \sum_{k=1}^{K} \rho_{\tau_k}\{y_i - z_i \hat{\beta}_L(\tau_k) - \hat{\alpha}_L(\tau_k)\} + \sum_{x_i > c} \sum_{k=1}^{K} \rho_{\tau_k}\{y_i - z_i \hat{\beta}_R(\tau_k) - \hat{\alpha}_R(\tau_k)\}, \quad (6)$$

where $\rho_\tau(u) = \{\tau - I(u < 0)\}u$ is the quantile loss function, $(\hat{\alpha}_L(\tau_k), \hat{\beta}_L(\tau_k))$ and $(\hat{\alpha}_R(\tau_k), \hat{\beta}_R(\tau_k))$ are estimated quantile regression coefficients in the left and right child nodes, respectively.

In practice, we consider grid search among sample quantiles for continuous split variables, and this method can be extended readily for categorical covariates to estimate the split set. By considering the composite loss across different quantile levels, the proposed method often leads to more robust and efficient estimation, especially in the case of heavy-tailed distributions.

Remark 2. The GQ method estimates the split point by minimizing the composite loss function in (6), assuming that the intercept and slope are split by the same threshold. Note that our primary interest is in the threshold splitting the slope (treatment effect). To accommodate cases where the thresholds vary in the intercept and slope, we could consider an alternative GQ_δ estimator of the desired split point, defined as

$$\hat{t} = \underset{c}{\operatorname{argmax}} \sum_{x_i \le c} \sum_{k=1}^{K} \left[\frac{\hat{\delta}(\tau_k)}{\sqrt{\operatorname{var}\{\hat{\beta}_L(\tau_k)\} + \operatorname{var}\{\hat{\beta}_R(\tau_k)\}}} \right]^2,$$

where $\hat{\beta}_L(\tau_k)$ and $\hat{\beta}_R(\tau_k)$ are estimated quantile regression slopes (coefficients associated with the treatment variable Z) in the left and right child nodes, respectively, and $\hat{\delta}(\tau_k) = \hat{\beta}_L(\tau_k) - \hat{\beta}_R(\tau_k)$ is the estimated treatment effect difference between the two child nodes. Even though the GQ_δ method can accommodate more general cases, the GQ method is computationally more efficient since it does not require variance estimation at any candidate split point. In addition, our numerical study shows that even when the thresholds due to the intercept and slope are different, the GQ method still performs competitively well or better if the differential effect in the slope is larger than that in the intercept. Therefore, throughout the rest of the article, we focus on the GQ approach for the split point estimation.

2.4. Selection and Confirmation of Predictive Subgroups

Evaluation of treatment efficacy in a given subgroup is important since it determines whether the subgroup should be selected or confirmed. Most existing methods focus on mean treatment effect and the two sample t-test or its regression counterpart is commonly used to measure the difference. But when two distribution functions differ only in the upper (or lower) tail, the mean-based tests may suffer from low power. Moreover, mean-based evaluations are easily affected by outliers and skewed distributions. To remedy such limitations, we propose to use the generalized quantile treatment effect (GQTE) testing procedure for the selection and confirmation of predictive subgroups.

2.4.1. Generalized Quantile Treatment Effect Test

Suppose that model (2) holds without covariate x_i ($\boldsymbol{\eta}(\tau) = \boldsymbol{\delta}(\tau) = \mathbf{0}$), and in this case $\beta(\tau)$ refers to the treatment effect at the τth quantile. We define the GQTE as

$$\theta = \int_0^1 \beta(\tau)\omega(\tau)d\tau,$$

where $\omega(\tau)$ is some weight function on $(0, 1)$. Replacing $\beta(\tau)$ with the sample counterpart, we can obtain the following estimator $\hat{\theta} = \int_0^1 \hat{\beta}(\tau)\omega(\tau)d\tau$. Assuming independence in the

control and treatment groups, the test statistic for assessing GQTE is given by

$$\mathcal{T}_n = \frac{\hat{\theta}}{\sqrt{\sigma^2(\omega, F_1)/n_1 + \sigma^2(\omega, F_0)/n_0}}.$$

Here n_z and F_z are the sample size and the response distribution for the treatment group $Z = z$, and

$$\sigma^2(\omega, F) = \int_0^1 \int_0^1 (s \wedge t - ts)[f\{F^{-1}(t)\} \times f\{F^{-1}(s)\}]^{-1} \omega(t)\omega(s)dsdt.$$

Under $H_0 : \theta = 0$, $\mathcal{T}_n$ converges to a standard normal distribution (Koenker and Portnoy 1987). By choosing different weight functions, the GQTE test is flexible and adaptive to different types of treatment effects. In particular, for the uniform weight $\omega(\tau) \equiv 1$, $\mathcal{T}_n$ reduces to the standard two sample t-test statistic.

The density f in $\sigma^2(\omega, F)$ is difficult to estimate. In practice, we consider using bootstrap to estimate the standard deviation of $\hat{\theta}$. Specifically, for each bootstrapped sample, we calculate the GQTE estimator $\hat{\theta}$, and then replace the denominator of $\mathcal{T}_n$ by the bootstrap sample standard deviation of $\hat{\theta}$.

2.4.2. Recommendation for Score and Weight Functions

In the proposed GQ algorithm, the rank score test is applied to determine the split variable at each node and the GQTE test is employed for selection and confirmation of predictive subgroups. The choice of scores in the former may affect the resulting tree structure, and the choice of weight functions in the latter may affect the assessment of treatment effect in a given subgroup and thus, the final subgroup identification. To obtain a consistent and reliable result, an appropriate pair of score and weight functions should be considered. We discussed in Section 2.2 how to choose the score function to optimize the power of the rank score test under local alternatives. For the GQTE test, we can choose the weight function to minimize the asymptotic variance of $\hat{\theta}$. The following proposition gives the optimal weight functions for the class of location-scale-shift models.

Proposition 2. Suppose that Model (2) holds without covariate X, and the error distribution F has a strictly positive and continuous density f. Then

(i) for location-shift models with $\beta(\tau) = \beta_0$, the optimal weight function takes the form

$$\omega(u) \propto \frac{(f')^2 - f''f}{f^2}\{F^{-1}(u)\}; \quad (7)$$

(ii) for scale-shift models with $\beta(\tau) = \beta_0 F^{-1}(\tau)$, the optimal weight function takes the form

$$\omega(u) \propto -\frac{f'}{f}\{F^{-1}(u)\} + F^{-1}(u)\frac{(f')^2 - f''f}{f^2}\{F^{-1}(u)\}. \quad (8)$$

Remark 3. With the optimal choice of weight function, the asymptotic variance of $\hat{\theta}$ reaches the Cramér–Rao bound. Below we summarize the optimal weight function for several scenarios.

(i) Location-shift model with Normal error: $\omega_{LN}(u) \equiv 1$, referred to as the Location-Normal weight.
(ii) Location-shift model with Logistic error: $\omega_{LL}(u) = u - u^2$, referred to as the Location-Logistic weight.
(iii) Scale-shift model with Normal error: $\omega_{SN}(u) = \Phi^{-1}(u)$, referred to as the Scale-Normal weight.
(iv) Scale-shit model with W error: $\omega_{SW}(u) = (1-u)^3/u$, referred to as the Scale-W weight.

To accommodate signals on the tails, we suggest trimmed weights such as the half Scale-Normal weight $\omega_{HSN}(u) = \Phi^{-1}(u)I(0.5 < u < 1)$ and the trimmed Scale-W weight $\omega_{TSW}(u) = (1-u)^3/uI(0.6 < u < 0.95)$. In practice it is often hard to determine the model and error distribution. Similar to the split variable selection, we also suggest an adaptive approach for applying the GQTE test. That is, we use the smallest p-value from the GQTE test based on multiple weight functions with Bonferroni correction to assess the significance of treatment effect of different forms. In our implementation, we consider four weight functions, $\omega_{LN}, \omega_{LL}, \omega_{HSN}$ and ω_{TSW}.

2.4.3. Confirmation of Predictive Subgroups

We propose to select candidate subgroups by assessing the significance of the GQTE for each node separately. However, it is known that inference on the subgroup identified from the same data may suffer from inflated false discovery rate (the chance of incorrectly identifying admissible subgroups when there's no predictive effects in any subsets); see, for example, Ruberg and Shen (2015) and Thomas and Bornkamp (2017). Therefore, the selected promising subgroups from the same sample are likely to be false positive due to over-optimism. To overcome this issue, we conduct analysis by dividing the entire data into training and testing samples. Let L, N_p and N_c be the maximum depth of the tree, the minimum sample size of a parent node and the minimum treatment group sample size, respectively. Below we provide a detailed description of the proposed algorithm:

(i) Divide the entire data into training and testing samples.
(ii) Refer to a node M at level l ($l = 0$ corresponds to the entire training sample) as the "parent group," and denote the sample size of M by n_p. For $l < L$, consider the following:
   - If $n_p < N_p$, then stop generating subsequent nodes and M is declared as a terminal node.
   - If $n_p \geq N_p$, divide M into two child nodes by (iia) selecting the split variable using the quantile rank score test as described in Section 2.2; (iib) estimating the associated split point based on Equation (6) in Section 2.3. We add an additional step to avoid searching the change point near the boundary, which may result in small sample size in one child node. In step (iib), if $\min(n_0^L, n_1^L, n_0^R, n_1^R) < N_c$, we exclude the corresponding candidate split point from consideration, where $n_0^L, n_1^L(n_0^R, n_1^R)$ are the sample sizes of the two treatment groups in the left (right) child node.

   If child nodes are generated, repeat (ii) with $l = l + 1$.
(iii) Select nodes with significant GQTEs (evaluated with the training sample) as the candidate subgroups.

(iv) Confirm the candidate subgroups by reevaluating the treatment effect using the testing-sample.

Remark 4. For the selection and confirmation of predictive subgroups in (iii) and (iv), we examine all the terminal nodes and the internal nodes (including the root node). Since our main target is to identify subgroups with enhanced treatment effect, if the root node is confirmed by the testing sample and it has the most significant GQTE among all confirmed subgroups, then there is no advantage to divide the entire data into subsamples and thus, no subgroup is identified. Our numerical results show that the proposed procedure provides adequate control of false subgroup discovery rate; see Section 3.3.

3. Simulation Study

We carry out simulations to assess the performance of the proposed GQ method on split variable selection, split point estimation and subgroup identification. For comparison, we also include the following methods, the interaction trees (IT) (Su et al. 2009), subgroup identification based on differential effect search (SIDES) (Lipkovich et al. 2011), model-based recursive partitioning (MOB) (Zeileis, Hothorn, and Hornik 2008; Seibold, Zeileis, and Hothorn 2016) and generalized unbiased interaction detection and estimation (GUIDE) (Loh 2002; Loh, He, and Man 2015). The IT method is applied using the R functions provided by the authors. The SIDES and MOB methods are based on the R packages *SIDES* and *partykit*, respectively. In our implementation of the MOB algorithm, at each node, we fit a linear model including all candidate split variables and the treatment indicator, with categorical covariates converted into dummy variables. Results from GUIDE are based on the software from *http://pages.stat.wisc.edu/~loh/guide.html* using the recommended "Gi" option with a constant model fitted at each node.

3.1. Split Variable Selection

We first compare the methods in terms of accuracy for selecting the true split variables in five simulation models, presented in Table 1. In all models, there are four candidate split variables X_1 –X_4 that are mutually independent and defined as:

$$X_1 \sim C(2), \quad X_2 \sim C(6), \quad X_3 \sim N(0,1), \quad X_4 \sim \exp(1), \tag{9}$$

where X_1 and X_2 are two categorical variables with two and six levels, respectively. Model 1 is the null model with prognostic effect (X_1) only. Models 2 and 3 are location-shift models with predictive variables X_1 and X_3, respectively. To access the performance of different methods under heterogeneity, we further consider two location-scale-shift models, and include a case where the errors follow a heavy-tailed t_2 distribution.

We generate 1000 observations for each example, and repeat the procedure 500 times. For our proposed method, we consider Normal, Wilcoxon, half Normal scale and trimmed Wilcoxon scores, referred to as GQ_N, GQ_W, GQ_{HNS} and GQ_{TW}, respectively. We also include the results given by the adaptive approach (GQ_A), assuming no prior information on either model type or error distribution.

Table 2 summarizes the selection percentages of all candidate split variables under the null Model 1. Since there's no predictive effect, a method has unbiased variable selection if the proportions are all around 25%. Both IT and SIDES employ exhaustive search algorithms to find the split variable, and they give biased results and tend to choose covariates that allow more splits. The MOB method is also biased toward X_2 that has six levels. In contrast, GUIDE and our proposed methods lead to selections with much smaller bias. Among different variations of the proposed method, the method with Wilcoxon score gives the smallest selection bias, and the adaptive method performs competitively well.

Selection percentages under the alternative Models 2–5 are reported in Table 3. Under the homogeneous Model 2, the proposed methods (specifically with Normal, Wilcoxon and adaptive scores) and GUIDE outperform the others. For Model 3 where the true split variable is X_3, IT gives the best selection accuracy. In Models 4 and 5, the errors depend on the split variable X_3, and $\boldsymbol{\delta}(\tau)$, and the interaction effect between Z and X_3 on the τth quantile of Y is increasing in τ. In these two models, the proposed method with the half Normal and trimmed Wilcoxon scores perform the best, which is not surprising since the signal is stronger on the right tail. Compared to existing approaches, the GQ methods show more advantages in cases with heteroscedastic or heavy-tailed errors. Across all different scenarios considered, the proposed adaptive method GQ_A is competitive to the best performer and thus, is recommended

Table 1. Simulation design for split variable selection.

Model	Error Dist.
1: $Y = 1 + I(X_1 = 1) + \epsilon$	$N(0,1)$
2: $Y = 1 + 0.4I(X_1 = 1)Z + \epsilon$	$N(0,1)$
3: $Y = 1 + 0.4I(X_3 > 0.5)Z + \epsilon$	$N(0,1)$
4: $Y = 1 + 0.4I(X_3 > 0.5)Z + (0.4I(X_3 > 0.5)Z + 1)\epsilon$	$N(0,1)$
5: $Y = 1 + 0.4I(X_3 > 0.5)Z + (0.4I(X_3 > 0.5)Z + 1)\epsilon/2$	t_2

Table 2. Percentage of times each covariate is selected as the splitting variable in the null model.

Var.	IT	SIDES	GQ_A	GQ_N	GQ_{HNS}	GQ_W	GQ_{TW}	MOB	GUIDE
X_1	2.4	5.0	29.2	28.6	27.2	26.4	27.0	12.6	27.4
X_2	21.4	34.6	22.6	23.4	20.0	24.4	25.6	53.6	23.6
X_3	40.2	30.0	26.0	21.8	31.6	23.4	27.4	18.8	22.8
X_4	36.0	30.4	22.2	26.2	21.2	25.8	20.0	15.0	26.2

Table 3. Percentages of times each covariate is selected as the splitting variable under alternative models.

Model	Var.	IT	SIDES	GQ_A	GQ_N	GQ_{HNS}	GQ_W	GQ_{TW}	MOB	GUIDE
2	X_1	65.4	76.8	90.8	92.8	76.6	92.0	84.4	77.8	92.0
	X_2	8.2	11.0	2.4	1.2	7.4	2.4	4.8	14.8	1.8
	X_3	13.4	5.8	4.8	4.2	9.0	3.8	7.2	3.6	4.6
	X_4	13.0	6.4	2.0	1.8	7.0	1.8	3.6	3.8	1.6
3	X_1	0.4	2.6	11.8	9.0	18.0	9.8	14.8	3.2	8.8
	X_2	4.8	11.0	7.4	7.0	12.8	7.8	10.6	13.0	7.4
	X_3	85.6	77.8	72.0	76.8	57.2	76.2	65.0	79.6	77.0
	X_4	9.2	8.6	8.8	7.2	12.0	6.2	9.6	4.2	6.8
4	X_1	1.0	3.4	4.6	13.0	2.8	13.2	6.2	5.8	8.6
	X_2	5.8	18.2	1.2	11.0	1.4	11.0	3.0	17.2	5.2
	X_3	81.4	65.2	92.0	66.8	93.2	65.6	87.6	72.0	80.6
	X_4	11.8	13.2	2.2	9.2	2.6	10.2	3.2	5.0	5.6
5	X_1	1.6	3.2	2.2	5.4	7.2	3.6	1.8	5.6	17.4
	X_2	12.8	20.6	2.2	5.2	6.8	2.6	2.0	33.8	14.4
	X_3	64.6	57.6	94.2	84.8	79.4	90.2	95.0	54.0	54.0
	X_4	21.0	18.6	1.4	4.6	6.6	3.6	1.2	6.6	14.2

NOTE: The true splitting variable is X_1 in Model 2 and X_3 in Models 3–5.

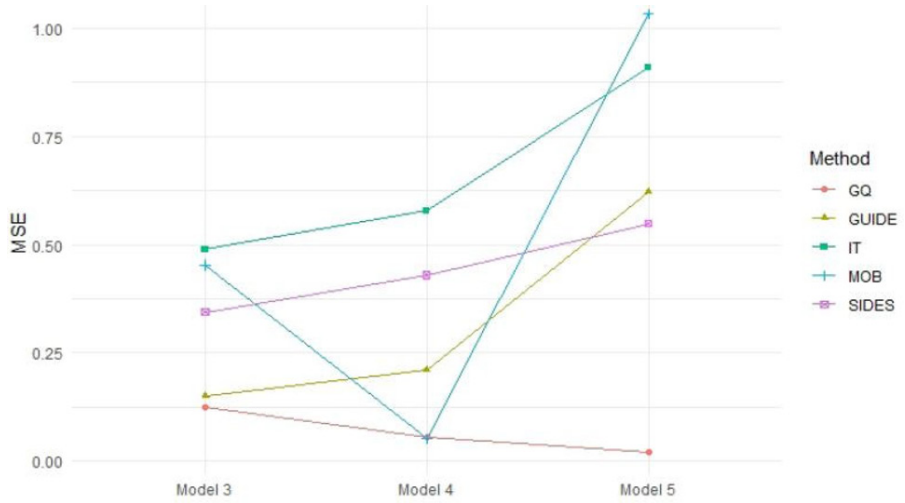

Figure 2. The mean squared errors (MSEs) of the estimated split points from different methods in Models 3–5.

Table 4. Simulation models considered for subgroup identification analysis, where $\epsilon \sim t_2$ in Model D, and $\epsilon \sim N(0,1)$ elsewhere.

Model	Formula
A	$Y = 1 + 4I(X_3 > 0.5) + \epsilon$
B	$Y = 1 + 4I(X_1 = 1) + \epsilon$
C	$Y = 1 + I(X_3 > 0.5) + I(X_4 > 0.5) + I(X_3 > 0.5)I(X_4 > 0.5)Z + \epsilon$
D	$Y = 1 + I(X_3 > 0.5) + I(X_4 > 0.5) + I(X_3 > 0.5)I(X_4 > 0.5)Z + \epsilon/2$
E	$Y = 1 + I(X_3 > 0.5) + I(X_3 > 0.5)Z + I(X_4 > 0.5) + [1 + 2I(X_3 > 0.5)Z]\epsilon$
F	$Y = 1 + I(X_3 > 0) + I(X_4 > 0.5) + [1 + 1.5I(X_3 > 0)I(\epsilon > 0.5)Z]\epsilon$

for practice when no prior information about the model and distribution is available.

3.2. Split Point Estimation

To assess the performance of different methods for split point estimation, we focus on Models 3–5, assuming that the split variable X_3 is correctly selected from the previous step. Figure 2 presents the mean squared errors of the estimated split points from different methods. Results show that our proposed method provides more efficient estimator than existing approaches, especially for cases with heteroscedastic or/and heavy-tailed errors.

3.3. Subgroup Identification Analysis

To compare the overall performance of different methods for subgroup identification, we consider six simulation models with various types of treatment effect and error distribution; see specifications in Table 4. For each model, we consider covariates $X_1 - X_4$ as candidate split variables. Models A and B correspond to the null model with no interactions between any covariates and the treatment indicator. Models C and D are location-shift models with the true subgroup $\{X_3 > 0.5, X_4 > 0.5\}$, and Model D has heavy-tailed t_2 errors. Model E is a location-scale-shift model with the true subgroup $\{X_3 > 0.5\}$. We further include Model F, where the enhanced treatment effect in the subgroup $\{X_3 > 0\}$ appears only on the right tail and this mimics the pattern seen from the rheumatoid arthritis data in Figure 1 and the AIDS clinical trial data in Figure 4.

Table 5. False positive rates (in percentages) of different methods for identifying at least one admissible subgroup under the null Models A and B.

Model	GQ	IT	SIDES	MOB	GUIDE
A	4.0	5.0	4.7	85.3	100.0
B	4.3	2.3	4.3	0.7	0.3

For each case, a random sample of size $n = 500$ is generated for both treatment groups, and the same procedure is repeated 300 times. It has been shown in Section 3.1 that the proposed split variable selection method with the adaptive score gives competitive results irregardless of the model and error distribution. Therefore, for the remaining of this simulation study, we will only focus on the proposed GQ method with the adaptive score and weight functions.

Table 5 summarizes the false positive rates for identifying admissible subgroups under the null Models A and B, defined as the proportion of times at least one subgroup is confirmed for SIDES and GQ and a nontrivial tree is generated after pruning for IT, GUIDE and MOB methods. Results show that the false positive rates of GQ, IT and SIDES are well controlled under the 5% nominal level. However, MOB and GUIDE have inflated false discovery rate (over 80%) when there exists a continuous prognostic variable. One possible reason is that both methods search through all distinct values to determine the best split point for a given split variable while there are infinite number of possible splits for a continuous prognostic variable.

Next, we compare the performance of different methods in identifying predictive subgroups under the alternative Models C–F. Let S_T be the true predictive subgroup, that is, the region with enhanced treatment effect, and let S_I denote the identified subgroup. For the GQ approach, S_I is chosen as the one that has the most significant generalized quantile treatment effect, that is, the smallest p-value with adaptive weights among all confirmed subgroups. For the other methods, S_I is defined as the leaf node with the most significant mean treatment difference. We use the following criteria to evaluate the performance for subgroup identification,

Table 6. Summaries of subgroup identification results of different methods under the alternative Models C–F.

Model	Method	PMR	Power	STY	SFY	R_{SS}	R_{MTE}
C	GQ	0.58	0.80	0.87	0.97	1.06	0.97
	IT	0.29	0.57	0.80	0.97	1.14	0.98
	SIDES	0.39	0.80	0.80	0.96	1.13	0.99
	MOB	0.86	0.97	0.95	1.00	1.08	0.97
	GUIDE	0.93	1.00	0.96	0.99	0.98	1.00
D	GQ	0.64	1.00	0.84	0.97	1.12	0.88
	IT	0.14	0.33	0.76	0.96	1.16	0.90
	SIDES	0.42	0.80	0.82	0.94	0.98	0.92
	MOB	0.50	0.83	0.80	0.96	1.24	0.72
	GUIDE	0.43	0.75	0.81	0.98	1.27	0.84
E	GQ	0.99	1.00	0.98	0.96	0.92	1.06
	IT	0.61	0.66	0.97	0.94	0.89	1.16
	SIDES	0.31	0.50	0.84	0.84	0.74	1.47
	MOB	0.61	0.62	0.99	0.97	0.92	1.10
	GUIDE	0.96	1.00	0.97	0.95	0.91	1.22
F	GQ	0.89	0.91	0.99	0.94	0.92	1.07
	IT	0.17	0.19	0.96	0.90	0.88	1.12
	SIDES	0.21	0.36	0.89	0.72	0.71	1.50
	MOB	0.54	0.56	0.98	0.96	0.96	1.08
	GUIDE	0.92	1.00	0.94	0.82	0.77	1.24

- Partial match rate (PMR): proportion of times that S_I is covered by an expanded region of the truth, that is, $S_I \subset \{X_3 > 0.4, X_4 > 0.4\}$ (Models C & D), $\{X_3 > 0.4\}$ (Model E) or $\{X_3 > -0.1\}$ (Model F);
- Power: proportion of times at least one admissible subgroup is identified;
- Sensitivity (STY): average of $|S_I \cap S_T|/|S_I|$, measuring the relative region that is identified correctly;
- Specificity (SFY): average of $|\bar{S}_I \cap \bar{S}_T|/|\bar{S}_I|$;
- R_{SS}: relative ratio of sizes, defined as the average of $|S_I|/|S_T|$;
- R_{MTE}: relative ratio of treatment effects, defined as the average of ratios between the mean treatment effect in S_I and S_T,

where $|\cdot|$ denotes the cardinality of a set. The last four criteria are calculated only among simulations when a subgroup is identified. A better performed method is expected to give higher values for the first four criteria and values closer to one for the last two criteria.

Table 6 summarizes the subgroup identification results. Generally speaking, the IT and SIDES methods show less power and lower accuracy than the other methods. The MOB and GUIDE methods give better performance for Model C with homogeneous and normal errors, but they are less effective than GQ for cases with heteroscedastic or heavy-tailed errors (Models D–F). Across all scenarios considered, the proposed algorithm GQ produces identified subgroups very close to the truth, and its overall performance is competitive or better than the other four methods, especially for the heterogeneous Models E & F when the treatment effect differs mostly in the upper tail.

4. Analysis of the AIDS ACTG175 Data

We illustrate the merit of the proposed method by analyzing data from AIDS Clinical Trials Group Protocol 175 (ACTG175), which contains 2139 HIV-infected patients and is available in R package *speff2trial*. Study subjects are randomized to four treatment arms: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine, and ddI monotherapy (Hammer et al. 1996). Following Lu, Zhang, and Zeng (2013) and Tsiatis et al. (2008), we choose the shifted CD4 count at 20 ± 5 weeks post-baseline as the continuous response Y (defined as the CD4 count at 20 ± 5 weeks post-baseline minus the CD4 count at baseline), and consider 12 baseline covariates as candidate split variables; see variable descriptions in Table 7. Existing works have shown that the cocktail treatment ZDV + ddI and ddI monotherapy tend to be effective for increasing the CD4 count of HIV-infected patients (Saravolatz et al. 1996; Collier et al. 1993). For our analysis, we focus on the difference between ZDV + ddI and ddI alone treatments. The entire sample consists of $n = 1083$ patients, with $n_t = 522$ subjects in the "treatment" group (ZDV+ddI) and the rest $n_c = 561$ belong to the "control" group (ddI). The objective of our study is to identify the subgroups with potential enhanced treatment effects.

Table 7. Description of variables in the ACTG175 data.

Variable	Description
Y	The response variable
Z	Treatment (ZDV+ddI) versus control (ddI)
age	Age in years
wtkg	Weight in kilogram
karnof	Karnofsky score (scale of 0–100)
cd40	CD4 count (cells/mm^3) at baseline
cd80	CD8 count (cells/mm^3) at baseline
hemo	Hemophilia (yes/no)
homo	Homosexual activity (yes/no)
drugs	History of intravenous drug use (yes/no)
race	White versus non-white
gender	Male and female
str2	Antiretroviral history (naive/experienced)
symptom	Symptomatic status (asymptomatic/symptomatic)

Figure 3 plots the sample quantiles of Y obtained from the entire sample for two treatment groups. The preliminary results suggest that CD4 counts of nearly 40% of the patient population showed no or little progression from the baseline. Our primary interest is to search for subgroups with enhanced treatment effect on increasing the CD4 count. Therefore, it is appropriate to focus more on patients who respond to the treatments, corresponding to the upper tail of the response distribution. To accommodate the tail effect, in addition to the adaptive score and weight functions, we also apply the proposed procedure with half Normal scale score (ϕ_{HNS}) and half Scale-Normal weight (ω_{HSN}).

We consider five methods to analyze this data: MOB, IT, SIDES, GUIDE and the proposed GQ method. For this data, the GUIDE method does not identify any admissible subgroups. The MOB method selects an artificial subgroup that shows no interaction with the treatment variable at mean or any quantiles. For GQ, IT and SIDES methods, half of the entire sample is used as a training set to find promising subgroups and the other half serves as a testing set to confirm the candidate subgroups. To accommodate the randomness from data partition, we repeat the process 100 times and record the identified subgroups for each repetition. Across 100 partitions, the proposed GQ algorithm with the adaptive approach leads to confirmed subgroups 34 times, among which the subgroup {Homo: no} (consisting of $n = 358$ subjects) is identified 30 times. Furthermore, for the assessment of treatment effect in the identified subgroups,

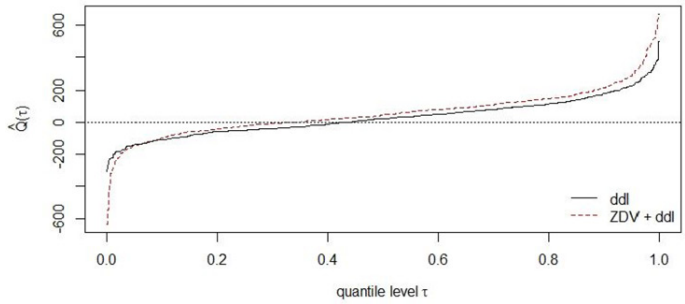

Figure 3. The sample quantiles of shifted CD4 count at 20 ± 5 weeks post-baseline for the ZDV+ddI and ddI treatment groups based on the entire sample.

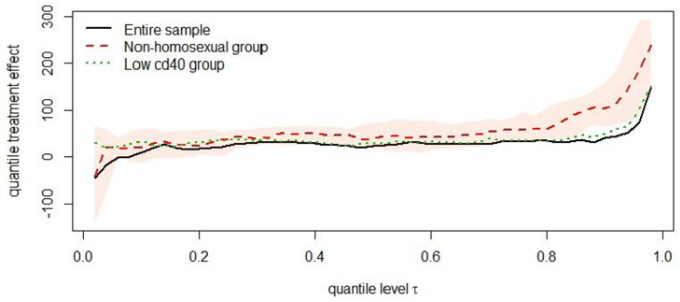

Figure 4. Estimated quantile treatment effect (ZDV+ddI against ddI) of the entire sample (solid), the nonhomosexual group (dashed) and the low cd40 group (dotted). The shaded area represents the 90% pointwise confidence band of the quantile treatment effect in the nonhomosexual group.

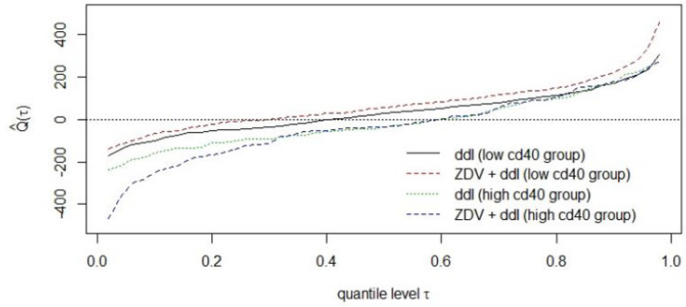

Figure 5. The sample quantiles of shifted CD4 count at 20 ± 5 weeks post-baseline for the ZDV+ddI and ddI treatment groups based on the low cd40 group and the high cd40 group.

the most significant GQTE test results are often times given by weight functions trimmed in upper tails. The GQ method with trimmed score (ϕ_{HNS}) and weight (ω_{HSN}) gives confirmed subgroups 53 times, and 47 times the identified subgroup is {Homo: no}, so the results agree quite well with those from the adaptive approach. The IT and SIDES methods identify admissible subgroups 22 and 75 times, respectively, and the most common subgroups identified are {cd40 $\leq$ 484} ($n = 959$, IT) and {cd40 $\leq$ 464} ($n = 922$, SIDES).

Since results from IT and SIDES are similar, we will next focus on comparing two subgroups: the low cd40 subgroup {cd40 $\leq$ 484} identified by the IT method, and the nonhomosexual subgroup {Homo: no} identified by the GQ method. We plot the estimated quantile treatment effect of the identified subgroups in Figure 4. The results suggest that the nonhomosexual group ({Homo: no}) exhibits an enhanced treatment effect, and the effect is more pronounced in the upper tail, especially for $\tau \in [0.8, 1)$. The treatment effect of the low cd40 subgroup ({cd40 $\leq$ 464}) differs from that of the entire sample only on the left tail for $\tau < 0.15$, mostly corresponding to those subjects with negative CD4 changes post treatments. To gain a deeper understanding, we plot in Figure 5 the sample quantiles of Y for two treatment groups in the low cd40 group and its complement, the high cd40 group ({cd40 $>$ 464}). The plots show a reverse effect of ZDV+ddI against ddI monoterapy in the high cd40 group, which might be due to the deleterious effect of the cocktail treatment containing ddI as observed in previous studies such as Lacombe et al. (2005) and Negredo et al. (2004). In addition, a higher proportion of subjects (about 60%) had no CD4 increase post treatments in the high cd40 group, compared to about 30% in the low cd40 group. One possible reason is the imbalance of antiretroviral history between two groups, and specifically, the higher proportion of antiretroviral-naive patients in the high cd40 group. Therefore, the differences between the low and high cd40 groups in antiretroviral history and the reverse drug effect of nonresponders are likely the main reasons for the low cd40 subgroup to be identified. In contrast, Figure 4 suggests that the subgroup identified by the GQ methods is driven by the

enhanced treatment effect in the nonhomosexual subgroup over the entire sample across quantiles but primarily on the right tails, so it aligns better with the research objective.

5. Discussion

We have developed a generalized quantile tree method for subgroup identification. Our numerical studies show that the method can capture different forms of heterogeneous treatment effects, and it leads to more accurate subgroup identification than existing mean-based methods for cases with heterogeneous and heavy-tailed errors.

The proposed method consists of three key steps: (1) split variable selection via the quantile rank score test; (2) split point estimation based on a composite quantile loss; and (3) evaluation of treatment effects by the GQTE test. The proposed GQ algorithm targets experiment data with balanced designs. The algorithm can be extended to observational studies by, for example, including the confounding variables as additionally protected predictors in all three steps. Such an extension will require more theoretical and practical investigations.

Our simulation study shows that the adaptive method is competitive to the best performer across all scenarios considered, even though no information about the latent model or error distribution is used. However, the adaptive approach may result in different choices of the score and weight functions in data splitting and evaluating treatment effects at each node, causing challenges in interpreting the final identified subgroup. For real applications, we suggest using the adaptive approach to conduct preliminary analysis, and then applying the proposed algorithm by choosing a specific pair of score and weight functions based on the preliminary results and the research interest; see for example the application on ACTG175 data in Section 4.

To control the overall Type I error of identifying at least one subgroup when there is no predictive effect in any subpopulation, we divide the entire data into training and testing sets to remedy the over-fitting problem. Simulation shows that this method provides decent control of the false positive rate. The random data partition, however, may affect the subgroup identification result for finite samples. The proposed algorithm can be enhanced by adopting the tree pruning and cross-validation procedure of CART (Breiman et al. 1984), though at the cost of computational time. Recently, Fuentes, Casella, and Wells (2018), and Guo and He (2020) proposed valid inference methods to address the over-optimism issue for inference on the best subgroup selected from the same sample. Adapting these approaches to conduct formal inference on subgroups identified by the GQ method without data partition is an interesting topic for future investigation.

Supplementary materials

Technical proofs, the R code of the proposed GQ method and some additional simulation results are provided in the supplementary materials.

Acknowledgments

The authors gratefully acknowledge Dr. Xiaogang Su for sharing their R code.

Funding

This work is partly supported by the IR/D program and grant DMS-1712760 from the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Berger, J. O., Wang, X., and Shen, L. (2014), "A Bayesian Approach to Subgroup Identification," *Journal of Biopharmaceutical Statistics*, 24, 110–129. [824]

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Boca Raton, FL: CRC Press. [824,833]

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011), "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections," *Biostatistics*, 12, 270–282. [824]

Chen, G., Zhong, H., Belousov, A., and Devanarayan, V. (2015), "A Prim Approach to Predictive-Signature Development for Patient Stratification," *Statistics in Medicine*, 34, 317–342. [824]

Collier, A. C., Coombs, R. W., Fischl, M. A., Skolnik, P. R., Northfelt, D., Boutin, P., Hooper, C. J., Kaplan, L. D., Volberding, P. A., Davis, L. G., Henrard, D. R., Weller, S., and Corey, L. (1993), "Combination Therapy with Zidovudine and Didanosine Compared with Zidovudine Alone in HIV-1 Infection," *Annals of Internal Medicine*, 119, 786–793. [831]

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011), "Subgroup Identification from Randomized Clinical Trial Data," *Statistics in Medicine*, 30, 2867–2880. [824]

Fuentes, C., Casella, G., Wells, M. T. (2018), "Confidence Intervals for the Means of the Selected Populations," *Electronic Journal of Statistics*, 12, 58–79. [833]

Gu, X., Yin, G., and Lee, J. J. (2013), "Bayesian Two-step Lasso Strategy for Biomarker Selection in Personalized Medicine Development for Time-to-Event Endpoints," *Contemporary Clinical Trials*, 36, 642–650. [824]

Guo, X., and He, X. (2020), "Inference on Selected Subgroups in Clinical Trials," *Journal of the American Statistical Association*, 116, 1498–1506. [833]

Gutenbrunner, C., Jurečková, J., Koenker, R., and Portnoy, S. (1993), "Tests of Linear Hypotheses Based on Regression Rank Scores," *Journal of Nonparametric Statistics*, 2, 307–331. [825]

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., Meriganet, T. C. (1996), "A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with cd4 Cell Counts from 200 to 500 per Cubic Millimeter," *New England Journal of Medicine*, 335, 1081–1090. [831]

He, X., Hsu, Y.-H., and Hu, M. (2010), "Detection of Treatment Effects by Covariate-Adjusted Expected Shortfall," *The Annals of Applied Statistics*, 4, 2114–2125. [825]

Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L., and Devanarayan, V. (2017), "Patient Subgroup Identification for Clinical Drug Development," *Statistics in Medicine*, 36, 1414–1428. [824]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470. [824]

Italiano, A. (2011), "Prognostic or Predictive? It's Time to Get Back to Definitions," *Journal of Clinical Oncology*, 29, 4718. [825]

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., and Branson, M. (2011), "Bayesian Models for Subgroup Analysis in Clinical Trials," *Clinical Trials*, 8, 129–143. [824]

Keystone, E. C., Kavanaugh, A. F., Sharp, J. T., Tannenbaum, H., Hua, Y., Teoh, L. S., Fischkoff, S. A., and Chartash, E. K. (2004), "Radiographic, Clinical, and Functional Outcomes of Treatment with Adalimumab (a Human Anti–tumor Necrosis Factor Monoclonal Antibody) in Patients with Active Rheumatoid Arthritis Receiving Concomitant Methotrexate Therapy: A Randomized, Placebo-Controlled, 52-week Trial," *Arthritis & Rheumatism*, 50, 1400–1411. [824]

Kocherginsky, M., He, X., and Mu, Y. (2005), "Practical Confidence Intervals for Regression Quantiles," *Journal of Computational and Graphical Statistics*, 14, 41–55. [826]

Koenker, R. (2010), "Rank Tests for Heterogeneous Treatment Effects with Covariates," in *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*, pp. 134–142, Institute of Mathematical Statistics. [825,827]

Koenker, R., and Machado, J. A. (1999), "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310. [826]

Koenker, R., and Portnoy, S. (1987), "L-Estimation for Linear Models," *Journal of the American statistical Association*, 82, 851–857. [828]

Kremer, J. M., Genant, H. K., Moreland, L. W., Russell, A. S., Emery, P., Abud-Mendoza, C., Szechinski, J., Li, T., Ge, Z., Becker, J.-C., and Westhovens, R. (2006), "Effects of Abatacept in Patients with Methotrexate-Resistant Active Rheumatoid Arthritis: A Randomized Trial," *Annals of Internal Medicine*, 144, 865–876. [824]

Lacombe, K., Pacanowski, J., Meynard, J.-L., Trylesinski, A., and Girard, P.-M. (2005), "Risk Factors for cd4 Lymphopenia in Patients Treated with a Tenofovir/Didanosine High Dose-Containing Highly Active Antiretroviral Therapy Regimen," *Aids*, 19, 1107–1108. [832]

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011), "Subgroup Identification Based on Differential Effect Search—A Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations," *Statistics in Medicine*, 30, 2601–2621. [824,825,829]

Loh, W.-Y. (2002), "Regression Tress with Unbiased Variable Selection and Interaction Detection," *Statistica Sinica*, 12, 361–386. [825,829]

Loh, W.-Y., He, X., and Man, M. (2015), "A Regression Tree Approach to Identifying Subgroups with Differential Treatment Effects," *Statistics in Medicine*, 34, 1818–1833. [824,825,829]

Lu, W., Zhang, H. H., and Zeng, D. (2013), "Variable Selection for Optimal Treatment Decision," *Statistical Methods in Medical Research*, 22, 493–504. [831]

Mehta, S., Shelling, A., Muthukaruppan, A., Lasham, A., Blenkiron, C., Laking, G., and Print, C. (2010), "Predictive and Prognostic Molecular Markers for Cancer Medicine," *Therapeutic Advances in Medical Oncology*, 2, 125–148. [825]

Mologen (2018), "Final Analysis of Impulse Study Confirms Topline Data with Positive Subgroup Results," Press Release. [824]

Negredo, E., Moltó, J., Burger, D., Viciana, P., Ribera, E., Paredes, R., Juan, M., Ruiz, L., Puig, J., Pruvost, A., Grassi, J., Masmitjà, E., Clotetet, B. (2004), "Unexpected cd4 Cell Count Decline in Patients Receiving Didanosine and Tenofovir-Based Regimens Despite Undetectable Viral Load," *Aids*, 18, 459–463. [832]

Ruberg, S. J., and Shen, L. (2015), "Personalized Medicine: Four Perspectives of Tailored Medicine," *Statistics in Biopharmaceutical Research*, 7, 214–229. [828]

Saravolatz, L. D., Winslow, D. L., Collins, G., Hodges, J. S., Pettinelli, C., Stein, D. S., Markowitz, N., Reves, R., Loveless, M. O., Crane, L., Thompson,M. and Abrams, D. (1996), "Zidovudine Alone or in Combination with Didanosine or Zalcitabine in HIV-Infected Patients with the Acquired Immunodeficiency Syndrome or Fewer than 200 cd4 Cells per Cubic Millimeter," *New England Journal of Medicine*, 335, 1099–1106. [831]

Seibold, H., Zeileis, A., and Hothorn, T. (2016), "Model-Based Recursive Partitioning for Subgroup Analyses," *The International Journal of Biostatistics*, 12, 45–63. [824,825,829]

Sivaganesan, S., Laud, P. W., and Müller, P. (2011), "A Bayesian Subgroup Analysis with a Zero-Enriched Polya Urn Scheme," *Statistics in Medicine*, 30, 312–323. [824]

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009), "Subgroup Analysis via Recursive Partitioning," *Journal of Machine Learning Research*, 10, 141–158. [824,825,829]

Thomas, M., and Bornkamp, B. (2017), "Comparing Approaches to Treatment Effect Estimation for Subgroups in Clinical Trials," *Statistics in Biopharmaceutical Research*, 9, 160–171. [828]

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008), "Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: A Principled yet Flexible Approach," *Statistics in Medicine*, 27, 4658–4677. [831]

van der Heijde, D., Klareskog, L., Rodriguez-Valverde, V., Codreanu, C., Bolosiu, H., Melo-Gomes, J., Tornero-Molina, J., Wajdula, J., Pedersen, R., Fatenejad, S., and TEMPO Study Investigators. (2006), "Comparison of Etanercept and Methotrexate, Alone and Combined, in the Treatment of Rheumatoid Arthritis: Two-Year Clinical and Radiographic Results from the Tempo Study, a Double-Blind, Randomized Trial," *Arthritis & Rheumatism*, 54, 1063–1074. [825]

Wang, H. J., and Fygenson, M. (2009), "Inference for Censored Quantile Regression Models in Longitudinal Studies," *The Annals of Statistics*, 37, 756–781. [826]

Zeileis, A., Hothorn, T., and Hornik, K. (2008), "Model-Based Recursive Partitioning," *Journal of Computational and Graphical Statistics*, 17, 492–514. [825,829]

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L.-J. (2013), "Effectively Selecting a Target Population for a Future Comparative Study," *Journal of the American Statistical Association*, 108, 527–539. [824]

Zou, H., and Yuan, M. (2008), "Composite Quantile Regression and the Oracle Model Selection Theory," *The Annals of Statistics*, 36, 1108–1126. [827]