Published in final edited form as:
J Am Stat Assoc. 2023 ; 118(542): 898–912. doi:10.1080/01621459.2021.1957900.

Inference for High-Dimensional Censored Quantile Regression

Zhe Fei[1], **Qi Zheng**[2], **Hyokyoung G. Hong**[3], **Yi Li**[4]

[1]Department of Biostatistics, University of California, Los Angeles

[2]Department of Bioinformatics and Biostatistics, University of Louisville

[3]Department of Statistics and Probability, Michigan State University

[4]Department of Biostatistics, University of Michigan

Abstract

With the availability of high dimensional genetic biomarkers, it is of interest to identify heterogeneous effects of these predictors on patients' survival, along with proper statistical inference. Censored quantile regression has emerged as a powerful tool for detecting heterogeneous effects of covariates on survival outcomes. To our knowledge, there is little work available to draw inference on the effects of high dimensional predictors for censored quantile regression. This paper proposes a novel procedure to draw inference on all predictors within the framework of global censored quantile regression, which investigates covariate-response associations over an interval of quantile levels, instead of a few discrete values. The proposed estimator combines a sequence of low dimensional model estimates that are based on multi-sample splittings and variable selection. We show that, under some regularity conditions, the estimator is consistent and asymptotically follows a Gaussian process indexed by the quantile level. Simulation studies indicate that our procedure can properly quantify the uncertainty of the estimates in high dimensional settings. We apply our method to analyze the heterogeneous effects of SNPs residing in lung cancer pathways on patients' survival, using the Boston Lung Cancer Survivor Cohort, a cancer epidemiology study on the molecular mechanism of lung cancer.

Keywords

Conditional Quantiles; Fused-HDCQR; High Dimensional Predictors; Statistical Inference; Survival Analysis

1 Introduction

Lung cancer presents much heterogeneity in etiology (McKay et al., 2017; Dong et al., 2012; Huang et al., 2009), and some genetic variants may insert different impacts on different quantile levels of survival time. For example, in the Boston Lung Cancer Survivor Cohort (Christiani, 2017), a cancer epidemiology cohort of over 11,000 lung cancer cases enrolled in the Boston area since 1992, it was found that SNP AX.37793583 (rs115952579), along with age, gender, cancer stage and smoking status, had heterogeneous effects on different quantiles of survival time. A total of 674 patients in the study were genotyped, with the goal of identifying lung cancer survival-predictive SNPs. Target gene approaches, which focus on SNPs residing in cancer-related gene pathways, are appealing for increased statistical power in detecting significant SNPs (Moon et al., 2003; Risch and Plass, 2008; Ho et al., 2019), and the investigators have identified SNPs residing in 14 well-known lung cancer-related genes (Zhu et al., 2017; Korpanty et al., 2014; Yamamoto et al., 2008; Kelley et al., 2001). One goal was to investigate whether and how each SNP might play a different role among the high-risk (i.e. lower quantiles of overall survival) and low-risk (i.e. higher quantiles of overall survival) cancer survivors.

Quantile regression (QR) (Koenker and Bassett Jr, 1978) is a significant extension of classic linear regression. By permitting the effects of active variables to vary across quantile levels, quantile regression can naturally accommodate and examine the heterogeneous impacts of biomarkers on different segments of the response variable’s conditional distribution. As survival data are subject to censoring and may be incomplete, QR methods developed for complete data may be unsuitable. Efforts have been devoted to developing censored quantile regression (CQR) (Powell, 1986; Portnoy, 2003; Peng and Huang, 2008, among others), which has become a useful alternative strategy to traditional survival models, such as the Cox model and accelerated failure time model. QR has also been widely studied to accommodate high dimensional predictors. For example, Wang et al. (2012) dealt with variable selection using non-convex penalization; Zheng et al. (2013) proposed an adaptive penalized quantile regression estimator that can select the true sparse model with high probability; and Fan et al. (2014) studied the penalized quantile regression with a weighted L_1 penalty in an ultra-high dimensional setting. As to high dimensional CQR (HDCQR), He et al. (2013) provided a model-free variable screening procedure for ultra-high dimensional covariates, and Zheng et al. (2018) proposed a penalized HDCQR built upon a stochastic integral based estimating equation. However, most of the existing works in HDCQR were designed to select a subset of predictors and estimate the effects of the selected variables, instead of drawing inference on high dimensional predictors.

Progress in high dimensional inferences has been made for linear and non-linear models (Zhang and Zhang, 2014; Bühlmann et al., 2014; Javanmard and Montanari, 2014; Ning and Liu, 2017; Fei et al., 2019, among others). For example, Meinshausen et al. (2009) proposed to aggregate p-values from multi-sample splittings for high dimensional linear regression. Another line of works referred to as *post-selection inference* includes Berk et al. (2013), Lee et al. (2016), and Belloni et al. (2019), which recently provided post-selection inference at fixed quantiles for complete data. However, these methods may not handle censored outcomes. For censored median regression, Shows et al. (2010) provided sparse estimation and inference, but it cannot handle high dimensional data.

We propose to draw inference on high dimensional HDCQR based on a splitting and fusing scheme, termed Fused-HDCQR. Utilizing a variable selection procedure for HDCQR such as Zheng et al. (2018), our method operates partial regression followed by smoothing. Specifically, partial regression allows us to estimate the effect of each predictor, regardless whether it is chosen by variable selection or not. The fused estimator aggregates the estimates based on multiple data-splittings and variable selection, with a variance estimator derived by the functional delta method (Efron, 2014; Wager and Athey, 2018). To comprehensively assess the covariate effects on the survival distribution, we adopt a “global” quantile model (Zheng et al., 2015) with the quantile level being over an interval, instead of the local CQR that focuses only on a few pre-specified quantile levels. The global quantile model may indeed address the molecular mechanism of lung cancer, our motivating disease, that hypothesizes that some genetic variants may cause heterogeneous impacts on different but unspecified segments of survival distribution (McKay et al., 2017; Dong et al., 2012; Huang et al., 2009).

Our work presents several advantages. First, compared to high dimensional Cox models (Zhao and Li, 2012; Fang et al., 2017; Kong et al., 2018), the employed HDCQR stems from the accelerated failure time model (Wei, 1992) and offers straightforward interpretations (Hong et al., 2019). Second, utilizing the global conditional quantile regression, it uses various segments of the conditional survival distribution to improve the robustness of variable selection and capture global sparsity. Third, our splitting-and-averaging scheme avoids the technicalities of estimating the precision matrix by inverting the $p \times p$ Hessian matrix of the log likelihood, which is a major challenge for debiased-LASSO type methods (Zhang and Zhang, 2014; Van de Geer et al., 2014) and is even more so if we apply debiased-LASSO to the CQR setting. Finally, as opposed to post-selection inferences (Belloni et al., 2019, among others), Fused-HDCQR accounts for variations in model selection and draws inference for all of the predictors.

The rest of the paper is organized as follows. Section 2 introduces the method, and Section 3 details the asymptotic properties. Section 4 derives a non-parametric variance estimator, Section 5 conducts simulation studies, and Section 6 applies the proposed method to analyze the BLCSC data. The technical details, such as proofs and additional lemmas, are relegated to the online Supplementary Material.

2 Model and Method

2.1 High dimensional censored quantile regression

Let T and C denote the survival outcome and censoring time, respectively. We assume that C is independent of T given $\widetilde{\mathbf{Z}}$, a $(p-1)$-dimensional vector of covariates $(p > 1)$. Let $X = \min\{T, C\}$, $\Delta = 1\{T \le C\}$, and $\mathbf{Z} = \left(1, \widetilde{\mathbf{Z}}^{\mathrm{T}}\right)^{\mathrm{T}}$, where $1\{\cdot\}$ is the binary indicator function. The observed data, $D^{(n)} = \{(X_i, \Delta_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, are n i.i.d. copies of $(X, \Delta, \mathbf{Z})$. With $Y = \log T$, let $Q_Y(\tau|\mathbf{Z}) = \inf\{t : \mathrm{P}(Y \le t|\mathbf{Z}) \ge \tau\}$ be the τ-th conditional quantile of Y given $\mathbf{Z}$. A global censored quantile regression model stipulates

$$Q_Y(\tau \mid \mathbf{Z}) = \mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta}^*(\tau), \ \tau \in (0, 1), \tag{1}$$

where $\boldsymbol{\beta}^*(\tau)$ is a p-dimensional vector of coefficients at τ. We aim to draw inference on $\beta_j^*(\tau)$ for each $\tau \in (0, \tau_U]$ and for all $j \in \{1, \ldots, p\}$, where $0 < \tau_U < 1$ is an upper bound for estimable quantiles subject to identifiability constraint caused by censoring (Peng and Huang, 2008).

Let $N(t) = 1\{\log X \le t, \Delta = 1\}$, $\Lambda_T(t|\mathbf{Z}) = -\log(1 - \mathrm{P}(\log T \le t|\mathbf{Z}))$, and $H(u) = -\log(1 - u)$. Then, $M(t) = N(t) - \Lambda_T(t \wedge \log X|\mathbf{Z})$ is a Martingale process under model (1) (Fleming and

Harrington, 2011) and hence $E(M(t)|\mathbf{Z}) = 0$. We use $N_i(t)$ and $M_i(t)$, $i = 1, \ldots, n$, to denote the sample analogs of $N(t)$ and $M(t)$. Let $\theta_i(\tau) = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)$ and

$$\mathbf{U}_n(\boldsymbol{\beta}, \tau) = n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\left\{N_i(\theta_i(\tau)) - \int_0^{\tau} 1\{\log X_i \geq \theta_i(u)\}dH(u)\right\}.$$

We denote the expectation of $\mathbf{U}_n(\boldsymbol{\beta}, \tau)$ by $\mathbf{u}(\boldsymbol{\beta}, \tau)$.

The Martingale property implies $\mathbf{u}(\boldsymbol{\beta}^*, \tau) = 0$ with $\tau \in [0, \tau_U]$, entailing the estimating equation with $\tau \in (0, \tau_U]$:

$$n^{1/2}\mathbf{U}_n(\boldsymbol{\beta}, \tau) = n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_i\left\{N_i(\theta_i(\tau)) - \int_0^{\tau} 1\{\log X_i \geq \theta_i(u)\}dH(u)\right\} = 0. \qquad (2)$$

The stochastic integral in (2) naturally suggests sequential estimation with respect to τ. We define a grid of quantile values $\Gamma_m = \{\tau_0, \tau_1, \ldots, \tau_m\}$ to cover the interval $[\nu, \tau_U]$, where $\tau_0 = \nu$ and $\tau_m = \tau_U$. The assumption on the lower bound $\nu > 0$ is made to circumvent the singularity problem with CQR at $\tau = 0$, as detailed in assumption (A1). In practice, ν is chosen such that only a small proportion of observations are censored below the ν-th quantile.

Then, $\hat{\boldsymbol{\beta}}(\tau_k)$'s, the estimates of $\boldsymbol{\beta}(\tau_k)$'s, $\tau_k \in \Gamma_m$ can be sequentially obtained by solving

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_i\left(N_i(\theta_i(\tau_k)) - \sum_{r=0}^{k-1}\int_{\tau_r}^{\tau_{r+1}} 1\left\{\log X_i \geq \hat{\theta}_i(\tau_r)\right\}dH(u)\right) = 0,$$

where $\hat{\theta}_i(\tau_k) = \mathbf{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\tau_k)$. Due to the monotonicity of $\theta_i(\tau)$ in τ, $\hat{\boldsymbol{\beta}}(\tau_k)$ can be solved efficiently via L_1-minimization. And $\hat{\boldsymbol{\beta}}(\tau)$, $\tau \in [\nu, \tau_U]$ is defined as a right-continuous piece-wise constant function that only jumps at the grid points. It can be shown that $\hat{\boldsymbol{\beta}}(\tau)$ is uniformly consistent and converges weakly to a mean zero Gaussian process for $\tau \in [\nu, \tau_U]$ when $p = o(n)$. More importantly, $\hat{\boldsymbol{\beta}}(\tau)$ provides a comprehensive understanding of the covariate effects on the conditional survival distribution over the quantile interval $[\nu, \tau_U]$. We incorporate this sequential estimating procedure for low dimensional CQR estimation in our proposed method.

In addition, our method requires dimension reduction, which can be accomplished by existing methods, including the screening method proposed by He et al. (2013) and the penalized estimation and selection procedure developed by Zheng et al. (2018). Specifically, Zheng et al. (2018) incorporated an L_1 penalty into the stochastic integral based estimating equation in (2) to obtain an L-HDCQR estimator, which achieves a uniform convergence rate of $\sqrt{q\log(p \vee n)/n}$, and results in "sure screening" variable selection with high probability, where q is defined in condition (A4). Zheng et al. (2018) also proposed an AL-HDCQR estimator by employing the Adaptive Lasso penalties, which attains a uniform convergence rate of $\sqrt{q\log(n)/n}$ and selection consistency.

2.2 Fused-HDCQR estimator

Our proposed Fused-HDCQR procedure consists of multiple data splitting, selecting variables, fitting low dimensional CQRs with partitioned data, applying *append-and-estimate* to all predictors, and aggregating those estimates.

1. With the full data $D^{(n)}$, determine via cross-validation the tuning parameter(s) λ_n of $\mathcal{S}$, an HDCQR variable selection method.
2. Let B be a large positive number. For each $b = 1, \ldots, B$,
   - **i.** randomly split the data into equal halves D_1^b and D_2^b;
   - **ii.** on D_1^b, apply the selection procedure $\mathcal{S}$ with λ_n on $[\nu, \boldsymbol{\tau}_U]$, to select a subset of predictors, denoted by $\widehat{S}_{\lambda_n}^b$, or $\widehat{S}^b$ for short;
   - **iii.** on D_2^b, for each $j = 1, \ldots, p$, fit the partial CQR using the subset of covariates $\widehat{S}_{+j}^b = \left\{j\right\} \cup \widehat{S}^b$, and denote the estimator by $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)$, $\boldsymbol{\tau} \in [\nu, \boldsymbol{\tau}_U]$. $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)$ is a right-continuous piecewise-constant function that only jumps at the grid points at $\boldsymbol{\tau}_k \in \Gamma_m$;
   - **iv.** denote the entry in $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)$ corresponding to Z_j by $\tilde{\boldsymbol{\beta}}_j^b(\tau) = \left(\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)\right)_j$.
3. Fusing: the final estimator of $\beta_j^*(\tau)$, $\boldsymbol{\tau} \in [\nu, \boldsymbol{\tau}_U]$, $j = 1, \ldots, p$ is

$$\widehat{\beta}_j(\tau) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\beta}_j^b(\tau)\,. \tag{3}$$

***Remark* 1.—**We could select different tuning parameters for $\mathcal{S}$ in each data split, but with much added computation. Our numerical evidence seemed to suggest that a globally chosen λ_n work well.

***Remark* 2.—**Our procedure needs a variable selection procedure to reduce dimension. For example, L-HDCQR selects the subset $\{j \in \{2, \cdots, p\} : \max_k |\widehat{\gamma}_j(\tau_k)| > a_0,\ \tau_k \in \Gamma_m\}$, where $\widehat{\gamma}_j(\tau_k)$‘s are the L-HDCQR estimates, and $a_0 > 0$ is a predetermined threshold. We start j with 2 as the intercept term (corresponding to $j = 1$) is always included in the model. In regards to the choice of variable selection methods, based on our experience, we can adopt the screening method in He et al. (2013) for fast computation, use L-HDCQR for detecting any non-zero effects in the quantile interval $[\nu, \boldsymbol{\tau}_U]$, and choose AL-HDCQR if we opt to select fewer predictors.

***Remark* 3.—**With the censored outcomes, we have used the deviance residual to define the K-fold cross-validation criterion as in Zheng et al. (2018) and selected λ_n by minimizing it. Specifically, we partition the data to K folds, and let $\widehat{\boldsymbol{\beta}}_\lambda^{(-k)}(\tau)$ be the penalized estimate of $\boldsymbol{\beta}(\boldsymbol{\tau})$ using all of the data excluding the k-th fold with a tuning parameter λ and $\boldsymbol{\tau} \in [\nu, \boldsymbol{\tau}_U]$, where $k = 1, \ldots, K$. Under the global CQR model (1), we define the cross-validation error as

$$\text{CV Error}(\lambda) = \sum_{k=1}^{K} \sum_{i \in \text{ fold } k} \int_{\nu}^{\tau_U} |D_i[\widehat{\boldsymbol{\beta}}_{\lambda}^{(-k)}(\tau)]| d\tau, \tag{4}$$

where

$$D_i[\boldsymbol{\beta}(\tau)] = \text{sign}\{M_i(\boldsymbol{\beta}(\tau))\}\sqrt{-2M_i(\boldsymbol{\beta}(\tau)) + \Delta_i \log\{\Delta_i - M_i(\boldsymbol{\beta}(\tau))\}}$$

with $M_i(\boldsymbol{\beta}(\tau)) = N_i\left(\mathbf{Z}_i^T\boldsymbol{\beta}(\tau)\right) - \int_{\nu}^{\tau} 1\left\{\log X_i \geq N_i\left(\mathbf{Z}_i^T\boldsymbol{\beta}(u)\right)\right\} dH(u) - \nu$. Here, $H(u) = -\log(1 - u)$, $N_i(\cdot)$ is the counting process, and $M_i(\boldsymbol{\beta}(\tau))$ is the Martingale residual under model (1) (Zheng et al., 2018).

3 Theoretical Studies

3.1 Notation and regularity conditions

For any vector $\delta \in \mathbf{R}^p$ and a subset $S \subset \{1, \ldots, p\}$, denote by S^C the complementary set and define $\|\delta\|_{r,S} = \|\delta_S\|_r$, the l_r-norm of the sub-vector δ_S, in which $\delta_{jS} = \delta_j$ if $j \in S$ and $\delta_{jS} = 0$ if $j \notin S$. We set the following conditions.

(A1) There exists a quantile ν and some constant $c > 0$ such that

$$n^{-1} \sum_{i=1}^{n} 1\left\{\log C_i \leq \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\nu)\right\}(1 - \Delta_i) \leq cn^{-1/2}$$

holds for sufficiently large n.

(A2) (*Bounded observations*) $\|\mathbf{Z}\|_{\infty} \leq C_0$. Without loss of generality, we assume $C_0 = 1$. In addition, $E|\log X| < \infty$.

(A3) (*Bounded densities*) Let $F_T(t|\mathbf{Z}) = \mathrm{P}(\log T \leq t|\mathbf{Z})$, $\Lambda_T(t|\mathbf{Z}) = -\log(1 - F_T(t|\mathbf{Z}))$, $F(t|\mathbf{Z}) = \mathrm{P}(\log X \leq t|\mathbf{Z})$, and $G(t|\mathbf{Z}) = \mathrm{P}(\log X \leq t, \Delta = 1|\mathbf{Z})$. Also, define $f(t|\mathbf{Z}) = dF(t|\mathbf{Z})/dt$, and $g(t|\mathbf{Z}) = dG(t|\mathbf{Z})/dt$.

a. There exist constants $\underline{f}$, $\bar{f}$, $\underline{g}$ and $\bar{g}$ such that

$$\underline{f} \leq \inf_{\mathbf{z}, \tau \in [\nu, \tau_U]} f\left(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \leq \sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} f\left(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \leq \bar{f},$$

$$\underline{g} \leq \inf_{\mathbf{z}, \tau \in [\nu, \tau_U]} g\left(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \leq \sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} g\left(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \leq \bar{g}.$$

b. There exist constant $\kappa > 0$ and A such that $\forall |t| \leq \kappa$,

$$\sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} \left| f\left(\mathbf{z}^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + t \mid \mathbf{z}\right) - f\left(\mathbf{z}^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \right| \le A|t|,$$

$$\sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} \left| g\left(\mathbf{z}^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) + t \mid \mathbf{z}\right) - g\left(\mathbf{z}^{\mathrm{T}} \boldsymbol{\beta}^*(\tau) \mid \mathbf{z}\right) \right| \le A|t|.$$

(A4) (*Sparsity*) Assume $\log p = o(n^{1/2})$, let

$$S_\tau = \left\{j : \boldsymbol{\beta}_j^*(\tau) \ne 0\right\},\ S^* = \bigcup_{\tau \in [\nu, \tau_U]} S_\tau = \left\{j : \sup_{\tau \in [\nu, \tau_U]} \left|\boldsymbol{\beta}_j^*(\tau)\right| > 0\right\},\ \text{and } q = |S^*|.$$

Let $\hat{S}$ be the index set of covariates selected by $\mathcal{S}$ with a tuning parameter λ_n. There exist constants $0 \le c_1 < 1/3$, c_2, K_1, $K_2 > 0$ such that $q \le K_1 n^{c_1}$, $|\hat{S}| \le K_1 n^{c_1}$, and

$$\mathrm{P}\left(S^* \subseteq \hat{S}\right) \ge 1 - K_2 (p \vee n)^{-1 - c_2}.$$

(A5) Let $\tilde{\mu}(\tau) = E\left[1\left\{\log X > \mathbf{Z}^{\mathbf{T}} \boldsymbol{\beta}^*(\tau)\right\}\right]$. There exists a positive constant L, such that $\left|\boldsymbol{\beta}_j^*(\tau_1) - \boldsymbol{\beta}_j^*(\tau_2)\right| \le L|\tau_1 - \tau_2|$ and $|\tilde{\mu}(\tau_1) - \tilde{\mu}(\tau_2)| \le L|\tau_1 - \tau_2|$, for all $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in (\nu, \boldsymbol{\tau}_U]$ and $1 \le j \le p$.

(A6) (*Eigenvalues*) $\delta^{\mathrm{T}} E[\mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}}] \delta / \|\delta\|^2$ is bounded below and above by $\lambda_{\min}$ and $\lambda_{\max}$, respectively, over $\|\delta\|_0 \le K_1 n^{c_1}$, $\boldsymbol{\delta} \ne 0$, where $0 < \lambda_{\min} < \lambda_{\max}$. (*nonlinear impact*) $c_2 := \inf_{\|\delta\|_0 \le K_1 n^{c_1}, \delta \ne \mathbf{0}} E\left[\left(\mathbf{Z}_i^{\mathrm{T}} \delta\right)^2\right]^{3/2} / E\left[\left|\mathbf{Z}_i^{\mathrm{T}} \delta\right|^3\right] > 0.$

(A7) Γ_m is equally gridded with $\boldsymbol{\tau}_k - \boldsymbol{\tau}_{k-1} = \epsilon_n$ for $\boldsymbol{\tau}_k \in \Gamma_m$, $k = 1, \ldots, m$. The grid size satisfies $\epsilon_n = c_0 n^{-1}$ for some constant c_0.

Assumption (A1) requires that the number of censored observations below the ν-th quantile does not exceed $cn^{1/2}$, which is satisfied if the lower bound of the censoring time C's support is greater than 0 and seems reasonable in real applications. As recommended in Zheng et al. (2018), ν is chosen such that only a small proportion of the observed survival times below the ν-th quantile are censored. (A2) assumes that the covariates are uniformly bounded. As pointed out by Zheng et al. (2015), the global linear quantile regression model is most meaningful when the covariates are confined to a compact set to avoid crossing of the quantile functions. (A3) ensures the positiveness of $f(t|\mathbf{Z})$ between $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}^*(\nu)$ and $\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}^*(\boldsymbol{\tau}_U)$, which is essential for the identifiability of $\boldsymbol{\beta}^*(\boldsymbol{\tau})$ for $\boldsymbol{\tau} < \boldsymbol{\tau}_U$. (A4) restricts the order of data dimensions, as well as the sparsity of $\boldsymbol{\beta}^*(\boldsymbol{\tau})$, which is necessary for the convergence of the low dimensional estimator in (2) (Condition C4 in Wang et al. (2012)). (A4) also characterizes the "sure screening" property by S. The asymptotic property does not assess the variability of selection with a finite sample. For high dimensional inference, it is crucial to account for such variability (Fei et al., 2019). Specifically, several variable selection methods for high dimensional CQR satisfy the sure screening property in (A4) with additional mild conditions.

- L-HDCQR: by Corollary 4.1 of Zheng et al. (2018), a *Beta-min* condition is required in addition to the set of conditions in this paper. Explicitly, there exist constants C_1, $C_2 > 0$, such that

$$\inf_{j \in S^*} \sup_{\tau \in [\tau_L, \tau_U]} \left|\boldsymbol{\beta}_j^*(\tau)\right| > C_1 \exp(C_2 q \tau_U) \sqrt{q \log(p \vee n)/n} + L\sqrt{q}\epsilon_n .$$

- AL-HDCQR: by Corollary 4.2 of Zheng et al. (2018), AL-HDCQR achieves the stronger *selection consistency* property, which implies the sure screening property.
- Quantile-adaptive Screening: by Theorem 3.3 of He et al. (2013), with a proper threshold value in their technical conditions, the screening procedure achieves the sure screening property.

(A5) characterizes the smoothness of $\boldsymbol{\beta}^*(\boldsymbol{\tau})$. (A6) is analogous to the assumptions on the covariance structure in the high dimensional literature (Zhao and Yu, 2006; Belloni and Chernozhukov, 2011; Fan et al., 2014; Van de Geer et al., 2014). As an extension to Condition C4 in Peng and Huang (2008), it ensures the convergence of low dimensional CQR but with a diverging number of covariates. (A7) details the fineness of Γ_m, which renders an adequate approximation to the stochastic integration in (2).

3.2 Theoretical properties of Fused-HDCQR

We first extend the results in Peng and Huang (2008) from a fixed p to a *p-diverges-butless-than-n* case. They are novel and critical extensions since we allow the true model size $q = |S^*|$ to increase with n, while the selected $\hat{S}^b$'s in the fused procedure vary around S^*. Specifically, we assume a subset $S \subset \{1, \ldots, p\}$ in Theorems 1 and 2, where $|S| \le K_1 n^{c_1}$, $0 \le c_1 < 1/3$ and $K_1 > 0$. Let $\acute{\boldsymbol{\beta}}_S(\tau)$, $\boldsymbol{\tau} \in [\nu, \boldsymbol{\tau}_U]$ be the estimator from Peng and Huang (2008) of fitting the CQR with $\mathbf{Z}_S$ over the $\boldsymbol{\tau}$-grid Γ_m.

***Theorem* 1.**—(Consistency with a diverging number of predictors) Under Conditions (A1) – (A7) and given a subset $S \subset \{1, \cdots, p\}$ such that $S^* \subseteq S$ and $|S| \le K_1 n^{c_1}$, there exist positive constants ζ_1 and ζ_2 such that

$$\sup_{\nu \le \tau \le \tau_U} \|\acute{\boldsymbol{\beta}}_S(\tau) - \boldsymbol{\beta}^*(\tau)\| \le \zeta_1 \exp(\zeta_2)(K_1 n^{c_1 - 1} \log n)^{1/2}$$

with probability at least $1 - 20 c_0^{-2} K_1 n^{c_1 - 2}$.

***Remark* 4.**—From the proofs of Propositions 1 and 2, it can be seen that ζ_1 and ζ_2 do not depend on the choice of S or n. Thus, ζ_1 and ζ_2 are universal for all possible S satisfying $S^* \subseteq S$ and $|S| \le K_1 n^{c_1}$.

Next, we derive the weak convergence of $\acute{\boldsymbol{\beta}}_j$ for any $j \in S$.

***Theorem* 2.—**(Weak convergence with a diverging number of covariates) Suppose Conditions (A1) – (A7) hold. Given a $S \subset \{1, \cdots, p\}$ such that $S^* \subseteq S$ and $|S| \le K_1 n^{c_1}$, for any $j \in S$,

$$\sqrt{n}\left(\acute{\boldsymbol{\beta}}_j(\tau) - \boldsymbol{\beta}_j^*(\tau)\right)$$

converges weakly to a mean zero Gaussian process for $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$.

In high dimensional settings, the next theorem shows that the fused estimator enjoys desirable theoretical properties.

***Theorem* 3.—**Consider the Fused-HDCQR estimator in (3). Under assumptions (A1) – (A7), for any $j \in \{1, \dots, p\}$,

$$\sqrt{n}\left(\hat{\beta}_j(\tau) - \boldsymbol{\beta}_j^*(\tau)\right)$$

converges weakly to a mean zero Gaussian process for $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$.

Our framework enables us to obtain the joint distribution of K-dimensional estimated coefficients, where K is a finite number. Let $\mathscr{K}$ be the collection of the indices of K covariates of interest. We can show that the weak convergence result of $\acute{\boldsymbol{\beta}}_{\mathscr{K}}(\tau)$, a K-dimensional subvector of the oracle estimator, still holds for $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$, that is, $\sqrt{n}(\acute{\boldsymbol{\beta}}_{\mathscr{K}}(\tau) - \boldsymbol{\beta}^*_{\mathscr{K}}(\tau))$, $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$ converges to a K-dimensional Gaussian distribution at any $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$. We only need to replace $\acute{\boldsymbol{\beta}}_j(\tau)$ by $\acute{\boldsymbol{\beta}}_{\mathscr{K}}(\tau)$ in the proof of Theorem 2 in the Appendix and slightly modify the arguments accordingly. Consequently, the term I in the proof of Theorem 3 still converges weakly to a mean zero Gaussian distribution, while the norms of items II and III are still $o_p(1)$. Therefore, Theorem 3 still holds for any K-dimensional subvector of $\hat{\beta}_{\mathscr{K}}(\tau)$, i.e., $\sqrt{n}(\hat{\beta}_{\mathscr{K}}(\tau) - \boldsymbol{\beta}^*_{\mathscr{K}}(\tau))$ converges to a mean zero K-dimensional Gaussian distribution at any $\boldsymbol{\tau} \in [\boldsymbol{\nu}, \boldsymbol{\tau}_U]$.

As shown in the proof, the covariance function of $\hat{\beta}_j(\tau)$ depends on the unknown active set S^*, the unknown conditional density functions $f(t|\mathbf{Z})$ and $g(t|\mathbf{Z})$, and other unknown quantities. Thus, it is not calculable. The next section proposes an alternative model-free variance estimator based on functional delta method and multi-sample splitting properties (Efron, 2014; Fei et al., 2019).

4 A Variance Estimator via the Functional Delta Method

Let $J_{bi} \in \{0, 1\}$ be the indicator of whether i^{th} observation is in the b^{th} sub-sample D_2^b, and $\boldsymbol{J}_{.i} = \boldsymbol{B}^{-1}\sum_{b=1}^{B} \boldsymbol{J}_{bi}$. We define the re-sampling covariances between J_{bi} and $\tilde{\boldsymbol{\beta}}_j^b(\tau_k)$ at $\boldsymbol{\tau}_k \in \Gamma_m$ for each $i = 1, \dots, n$ as

$$\mathbf{s}_{ij}(\tau_k) = \frac{1}{B}\sum_{b=1}^{B}(J_{bi} - J_i)\left(\tilde{\boldsymbol{\beta}}_j^b(\tau_k) - \hat{\boldsymbol{\beta}}_j(\tau_k)\right);$$

$$\mathbf{S}_j(\tau_k) = (\mathbf{s}_{1j}(\tau_k), \mathbf{s}_{2j}(\tau_k), \ldots, \mathbf{s}_{nj}(\tau_k))^{\mathrm{T}}.$$

Let $n_1 = \left|D_2^b\right|$. The covariance between $\hat{\boldsymbol{\beta}}_j(\tau_k)$ and $\hat{\boldsymbol{\beta}}_j(\tau_\ell)$ is estimated by

$$\widehat{\mathrm{Cov}}_j(\tau_k, \tau_\ell) = \frac{n-1}{n}\left(\frac{n}{n-n_1}\right)^2\sum_{i=1}^{n}\mathbf{s}_{ij}(\tau_k)\mathbf{s}_{ij}(\tau_\ell) = \frac{n(n-1)}{(n-n_1)^2}\mathbf{S}_j^{\mathrm{T}}(\tau_k)\mathbf{S}_j(\tau_\ell),$$

where the multiplier $n(n-1)/(n-n_1)^2$ is a finite-sample correction for the sub-sampling (Wager and Athey, 2018). Thus a variance estimator for $\hat{\beta}_j(\tau_k)$ is

$$\hat{V}_j(\tau_k) = \frac{n(n-1)}{(n-n_1)^2}\mathbf{S}_j^{\mathrm{T}}(\tau_k)\mathbf{S}_j(\tau_k). \tag{5}$$

It is shown in Wager and Athey (2018) that (5) is consistent, i.e., $\hat{V}_j(\tau_k)/\mathrm{Var}\left(\hat{\beta}_j(\tau_k)\right) \xrightarrow{p} 1$ as $n,B \rightarrow \infty$. Furthermore, for a finite B, we propose a bias corrected version of (5):

$$\hat{V}_j^B(\tau_k) = \hat{V}_j(\tau_k) - \frac{nn_1}{B(n-n_1)}\left\{B^{-1}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}_j^b(\tau_k) - \hat{\boldsymbol{\beta}}_j(\tau_k)\right)^2\right\},\ \tau_k \in \Gamma_m. \tag{6}$$

The correction term in (6) is a suitable multiplier of the re-sampling variance of $\tilde{\boldsymbol{\beta}}_j^b(\tau_k)$'s, which converges to zero as $n \rightarrow \infty$ and $n_1 = O(n)$, and the two variance estimators in (5) and (6) are asymptotically equivalent. However, $\hat{V}_j(\tau_k)$ in (5) requires B to be of order $n^{3/2}$ to reduce the Monte Carlo noise below the sampling noise, while $\hat{V}_j^B(\tau_k)$ in (6) only requires B to be of order n to achieve the same (Wager et al., 2014).

Since $\hat{\boldsymbol{\beta}}_j(\tau)$ converges weakly to a Gaussian process by Theorem 3, and our variance estimators are consistent on the grid points, we define the asymptotic $100(1-\alpha)\%$ local confidence intervals for $\boldsymbol{\beta}_j^*(\tau_k)$ at any $\tau_k \in \Gamma_m$ as

$$\left(\hat{\boldsymbol{\beta}}_j(\tau_k) - \Phi^{-1}(1-\alpha/2)\sqrt{\hat{V}_j^B(\tau_k)},\ \hat{\boldsymbol{\beta}}_j(\tau_k) + \Phi^{-1}(1-\alpha/2)\sqrt{\hat{V}_j^B(\tau_k)}\right),$$

where $\hat{V}_j^B(\tau_k)$ is the variance estimator in (6), and Φ is the standard normal cumulative distribution function. The p-value of testing $H_0: \beta_j^*(\tau_k) = 0$ for each $\tau_k \in \Gamma_m$ is

$$2 \times \left\{1 - \Phi\left(\left|\hat{\boldsymbol{\beta}}_j(\tau_k)\right| / \sqrt{\hat{V}_j^B(\tau_k)}\right)\right\}.$$

5 Simulation Studies

In various settings, we have compared the proposed method, Fused-HDCQR (referred to as “Fused” in the tables and figures hereafter), with some competing methods in quantile regression or high dimensional inference. These methods include Wang et al. (2012) (“W12”) and Fan et al. (2014) (“F14”) for quantile regression; Zheng et al. (2018) (“Z18”) for censored quantile regression; and Meinshausen et al. (2009) (“M09”) for inference with aggregated p-values from multi-sample splittings.

In the simulations and the data analysis, we choose L-HDCQR described in Section 3 as the variable selection tool for Fused-HDCQR. We also explore the feasibility of using other alternatives for variable selection, such as Fan et al. (2009) (“F09”) and M09.

When implementing Fused-HDCQR, we specify the number of splits as $B = 300$, the quantile interval as $[\nu, \tau_U] = [0.1, 0.8]$, and the grid length as $m = n/\log p$. The tuning parameter is chosen by minimizing the 5-fold cross-validation error as in (4). We study the following examples with sparse non-zero effects, some of which are heterogeneous.

Example 1.

The event times are generated by

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + \varepsilon_i,\ i = 1, \ldots, n,$$

where the coefficient vector **b** are sparse with $b_{20} = 0.5$, $b_{40} = 1$, $b_{60} = 1.5$, $b_j = 0$ for all other j’s, and $\varepsilon_i \sim N(0, 1)$. Therefore, the true coefficients are $\boldsymbol{\beta}^*(\tau) = (Q_\varepsilon(\tau), \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}$ for all $\tau \in (0, 1)$, where $Q_\varepsilon(\tau)$, τ-th quantile of the distribution of ε, is the intercept. $\widetilde{Z}_{j,i}$‘s are i.i.d. Unif(−1, 1) for $j \in \{1, \ldots, p\}$. The censoring time is generated independently as $\log C_i = N(0, 16) + N(-5, 1) + N(8, 0.25)$, which gives a censoring rate around 25%.

Example 2.

The event times follow

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + 1.5\ \widetilde{Z}_{3,i}\varepsilon_i, \tag{7}$$

where $b_{20} = 1$, $b_{40} = 1.5$, $b_{60} = 2$, $b_j = 0$ for all other j’s, and $\varepsilon_i \sim N(0, 1)$. We first generate $\mathbf{Z}_i' \sim N_p(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{k\ell})_{p\times p}$, $\sigma_{k\ell} = 0.3^{|k-\ell|}$ the AR(1) correlation structure, and then let $\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i'$, except for the third covariate $\widetilde{Z}_{3,i} = |Z_{3,i}'| + 0.5$. Therefore $\boldsymbol{\beta}_1^*(\tau) = 0$, $\boldsymbol{\beta}_4^*(\tau) = 1.5Q_\varepsilon(\tau)$, and $\boldsymbol{\beta}_j^*(\tau) = b_{j+1}$, for all other j’s. The censoring time is generated independently as $\log C_i = N(0, 16) + N(-4, 1) + N(8, 0.25)$, which gives a censoring rate around 23%.

Example 3.

The event times follow

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + \phi_1(\xi_i)\ \widetilde{Z}_{1,i} + \phi_{10}(\xi_i)\ \widetilde{Z}_{4,i},$$

where $b_{20} = 1$, $b_{40} = 1.5$, $b_{60} = 2$, $b_j = 0$ for all other j's, $\xi_i \sim N(0, 1)$, and ϕ_1, ϕ_{10} are monotone functions as the dashed lines in Figure 1, both are continuous with zero and non-zero pieces over τ. We first generate $\mathbf{Z}_i^{\prime} \sim N_p(0, \mathbf{\Sigma})$ as in Example 2, and then let $\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i^{\prime}$, except $\widetilde{Z}_{1,i} = |Z_{1,i}^{\prime}| + 0.5$ and $\widetilde{Z}_{10,i} = |Z_{10,i}^{\prime}| + 0.5$. Therefore $\boldsymbol{\beta}_1^*(\tau) = 0$, $\boldsymbol{\beta}_2^*(\tau) = \phi_1(\tau)$, $\boldsymbol{\beta}_{11}^*(\tau) = \phi_{10}(\tau)$, and $\boldsymbol{\beta}_j^*(\tau) = b_{j+1}$, for all other j's. The censoring time is generated independently as $\log C_i = N(0, 16) + N(-4, 1) + N(10, 0.25)$, which gives a censoring rate around 20%.

For each of these examples, we set (n, p)=(300, 1000) and (700, 1000) to study the impacts of the sample size and the number of variables and how the methods fare when $p > n$. In Example 3, which mimics the real data example in Section 6 most closely, we have also explored (n, p) = (700, 2000), which is roughly equal to the dimension of the real dataset. For every parameter configuration, a total of 100 independent datasets are generated, and we report the averaged results from these replications, unless specified otherwise. The number of 100 is chosen because the penalized methods for high dimensional CQR are in general computationally intensive and take much computing time for one simulated dataset (Table 5).

We first evaluate the feasibility of using various variable selection tools for our proposed method. Comparisons of true positives and false negatives among F09, M09, and L-HDCQR under Examples 1–3 are reported in Table 1. F09 presents a subpar performance because, by taking intersections of variables selected from different partitions of data, it tends to miss out some true signals and thus have fewer true positives. In contrast, L-HDCQR retains more true positives than both F09 and M09, while having larger false positives. Because our method requires the variable selection step to include the true signals with high probability, even at the cost of some false positives, we have opted to use L-HDCQR as the screening tool for our method.

We next compare the performance of Fused-HDCQR with other high dimensional quantile regression methods at τ = .25, .5, .75 under Example 1. As a benchmark for comparisons, we also compute the oracle estimates based on the true model (with S^* known). As W12, F14, and Z18 provide coefficient estimates without standard errors (SEs), only the estimation biases are reported for them, while the average SEs, empirical standard deviations (SDs) and coverage probabilities of the confidence intervals are reported for our method. Table 2 shows that Fused-HDCQR presents the smallest biases, which are comparable to those of the oracle estimates. In contrast, Z18 has smaller biases when the sample size is large, and larger biases otherwise, while W12 and F14 incur substantial biases since they are not designed for censored data. Moreover, the SEs based on Fused-HDCQR agree with the empirical SDs of the estimates. The consistent estimates of coefficients and SEs obtained by Fused-HDCQR lead to proper coverage probabilities around the 0.95 nominal level. In addition, the coverage probabilities improve as n increases.

Table 2 also concerns the power for detection of signals. Since W12, F14, and Z18 cannot draw inference and, in general, there is lack of literature that deals with inference for HDCQR, we compare our method with the aggregated p-value approach (M09) in the quantile setting, though M09 originated from linear regression. The results indicate that

Fused-HDCQR outperforms M09, and presents adequate testing power when the effect size is moderate or large.

Table 3 summarizes the results from Example 2 with the heterogeneous effect β_4 varying with τ. We compare the estimation accuracy between Fused-HDCQR and Z18, as well as the statistical power between Fused-HDCQR and M09. Again, Fused-HDCQR presents smaller biases than Z18 and a higher power than M09. To assess whether the tuning parameters selected as in Remark 3 help the variable selection method (L-HDCQR) used by Fused-HDCQR satisfy assumption (A4) in Section 3, we report the selection frequency of each signal variable in Table 3 (and also in Table 4), and observe that the selection frequency increases as the sample size increases, hinting that assumption (A4) may be satisfied with these selected tuning parameters.

Table 4 summarizes the results based on Example 3. For the two heterogeneous effects β_2 and β_{11} that vary with τ, their estimation biases of Fused-HDCQR become smaller and the estimated SEs are closer to the empirical ones as n increases. Figure 1 shows that the Fused-HDCQR estimates agree with the oracle estimates and the truth, except at the change points, and have narrower confidence intervals with a larger n.

Finally, we compare the computation intensity among Z18, M09, W12, F14, and Fused-HDCQR under Example 1 and report in Table 5 the average computing time per dataset. Our method is the most computationally intensive, because it involves multiple data-splittings and draws inferences on all of the p coefficients. However, by utilizing parallel computing, we have managed to reduce the computational time to the same order of Z18, W12, and F14 that are based on penalized regression.

6 Application to the Boston Lung Cancer Survivor Study (BLCSC)

Detection of molecular profiles related to cancer patients' survival can aid personalized treatment, leading to prolonged survival and improved quality of life. In a subset of BLCSC samples, 674 lung cancer patients were measured with survival times, along with 40, 000 SNPs and clinical indicators, such as lung cancer subtypes (adenocarcinoma, squamous cell carcinoma, or others), cancer stages (1–4), age, gender, education level (≤ high school or > high school) and smoking status (active or non-active smokers); see Table 6 for patients' characteristics. The censoring rate was 23% and a total of 518 deaths were observed during the followup period, with the observed followup time varying from 13 to 8, 584 days.

We could have included all 40,000 SNPs in our analysis. However, for more statistical power, we opt for the targeted gene approach by focusing on 2,002 SNPs residing in 14 genes identified to be cancer related, namely, ALK, BRAF, BRCA1, EGFR, ERBB2, ERCC1, KRAS, MET, PIK3CA, RET, ROS1, RRM1, TP53, and TYMS (Brose et al., 2002; Toyooka et al., 2003; Paez et al., 2004; Soda et al., 2007). Pinpointing the effects of individual loci within the targeted genes is helpful for understanding disease mechanisms (Evans et al., 2011; D'Antonio et al., 2019) and designing gene therapies (Pâques and Duchateau, 2007; Hanawa et al., 2004). We also adjust for patients' clinical and environmental characteristics listed in Table 6, which gives a total of $p = 2,011$ predictors.

We apply Fused-HDCQR to compute the coefficient estimates (3) and variance estimates (6). We set the quantile interval to be [0.2, 0.7], which is wide enough to cover high risk and low risk groups and, in the meantime, ensures the quantile parameters be estimable in the presence of censoring (Zheng et al., 2015). We choose the lower bound $\tau_0 = \nu = 0.1$ to circumvent the singularity problem with CQR at $\tau = 0$, because few (< 2%) observations are censored below the ν-th quantile. With $\epsilon_n = 01$, we form the τ-grid Γ_m of length $m = 61$. We set $B = 750$ as the number of re-samples, which is sufficiently large based on our numerical experience. To determine the tuning parameter λ_n in L-HDCQR for selection, we use 5-fold cross-validation as specified in Remark 3.

For ease of presentation, we summarize the results evaluated at 6 quantile levels, $\tau = 0.2, 0.3, \ldots, 0.7$, instead of the whole grid Γ_m. To highlight the findings of the high risk group, we rank all SNPs based on their p-values at $\tau = 0.2$. After Bonferroni correction for multiple testing, there are 83 significant SNPs with the overall type I error of $\alpha = 0.05$. Our method estimates the coefficients and the p-values for *all* predictors, and we only present the results for the patient characteristics, the top 10 significant SNPs, and the 3 least significant SNPs in Figure 2 and Table 7. The estimated coefficient of active smoking drops from −0.42 ($p = 0.0011$) to −0.53 ($p = 0.0005$) as τ changes from 0.2 to 0.5, and then increases to −0.31 ($p = 0.038$) as τ changes to 0.7, suggesting that active smoking might be more harmful to the high or median risk groups than the low risk group of patients. The most significant SNP at $\tau = 0.2$ is AX.37793583 T, which remains significant throughout $\tau = 0.2$ to $\tau = 0.7$. However, its estimated coefficient decreases from 2.75 ($\tau = 0.2$) to 1.39 ($\tau = 0.7$), indicating its heterogeneous impacts on survival, i.e. stronger protective effect at lower quantiles and vice versa.

The effects of some SNPs are nearly zero for higher quantiles. For example, the estimated coefficient of AX.15207405 G decreases from 2.03 ($\tau = 0.2$; $p = 10^{-24}$) to −0.05 ($\tau = 0.7$; $p = 0.92$), with the estimated standard error increasing from 0.20 to 0.48. Similarly, the estimated coefficient of AX.40182999 A decreases from 1.5 ($\tau = 0.2$; $p = 9.6\times10^{-13}$) to −0.01 ($\tau = 0.7$; $p = 0.96$). The results again hint at heterogeneous SNP effects in various risk groups, which cannot be detected using traditional Cox models.

Finally, our results shed light on the roles of SNPs in the high risk group (i.e. lower quantiles). Specifically, we map the 83 SNPs with significant effects at the 0.2-th quantile by Fused-HDCQR to the corresponding genes and rank the genes by the number of significant SNPs (over total number of SNPs for each gene in the parenthesis), which are TP53 (14/321), RRM1 (14/174), ERCC1 (10/167), BRCA1 (10/114), ALK (8/163), ROS1 (5/294), EGFR (5/261), ERBB2 (4/167), and 6 other genes with numbers of significant SNPs less than 4. While these genes were reported to be associated with lung cancer (Toyooka et al., 2003; Takeuchi et al., 2012; Rosell et al., 2011; Lord et al., 2002; Zheng et al., 2007; Sasaki et al., 2006; Brose et al., 2002), our analysis provides more detailed information as to which SNPs and locations of the genes are jointly associated with the lung cancer survival, as well as the estimated effects and uncertainties. Analysis of heterogeneous SNP effects has been gaining increasing research attention in lung cancer research (McKay et al., 2017; Dong et al., 2012; Huang et al., 2009), and beyond it (Garcia-Closas et al., 2008; Cheng et al., 2010; Gulati et al., 2014).

7 Conclusions

Our proposed procedure involves repeated estimates from low dimensional CQRs, which are computationally straightforward and can be efficiently implemented with parallel computing. We require the variable selection to possess a sure screening property as in condition (A4). This seems to be supported by our simulations, which find our procedure works well when the variable selection method can select a superset of the true model with high probability. Our condition is much weaker than a stringent condition of selection consistency as specified in Fei et al. (2019).

In regards to the selection of B, we recommend B to be in the same order of the sample size n. Smaller B might not affect coefficient estimation much, but it would yield biased standard errors for inference. In addition, we opt to define Γ_m by setting the grid as $n/\log p$ equally spaced points between τ_0 and τ_U. This may cover the quantile interval well, with reasonable computation efficiency.

There are open questions left to be addressed. First, substantial work is needed when predictors are highly correlated as the performance of our method, like the other competing methods, deteriorates when correlations among predictors become stronger. Second, it is of interest to investigate an alternative method when the sparsity condition fails. For example, it is challenging to find an effective strategy to draw inference when a non-negligible portion of predictors have small but non-zero effects. We will pursue them elsewhere.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are deeply grateful toward the Editor, the AE and the two referees for their constructive comments and suggestions that have helped improve the manuscript substantially. We thank our long time collaborator, David Christiani, Harvard Medical School, for providing the Boston Lung Cancer Survivor Cohort data. The work is partially supported by a grant from NIH and NSF.

References

Belloni A and Chernozhukov V (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics 39(1), 82–130.

Belloni A, Chernozhukov V, and Kato K (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. Journal of the American Statistical Association 114(526), 749–758.

Berk R, Brown L, Buja A, Zhang K, and Zhao L (2013). Valid post-selection inference. The Annals of Statistics 41(2), 802–837.

Brose MS, Volpe P, Feldman M, Kumar M, Rishi I, Gerrero R, et al. (2002). BRAF and RAS mutations in human lung cancer and melanoma. Cancer research 62(23), 6997–7000. [PubMed: 12460918]

Bühlmann P, Kalisch M, and Meier L (2014). High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application 1, 255–278.

Cheng I, Plummer SJ, Neslund-Dudas C, Klein EA, Casey G, Rybicki BA, and Witte JS (2010). Prostate cancer susceptibility variants confer increased risk of disease progression. Cancer Epidemiology and Prevention Biomarkers 19(9), 2124–2132.

Christiani DC (2017). The Boston lung cancer survival cohort. http://grantome.com/grant/NIH/U01-CA209414-01A1. [Online; accessed November 27, 2018].

D'Antonio M, Reyna J, Jakubosky D, Donovan MK, Bonder M-J, et al. (2019). Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. eLife 8, e48476. [PubMed: 31746734]

Dong J, Hu Z, Shu Y, Pan S, Chen W, Wang Y, et al. (2012). Potentially functional polymorphisms in dna repair genes and non-small-cell lung cancer survival: A pathway-based analysis. Molecular carcinogenesis 51(7), 546–552. [PubMed: 21739480]

Efron B (2014). Estimation and accuracy after model selection. Journal of the American Statistical Association 109(507), 991–1007. [PubMed: 25346558]

Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, Kochan G, et al. (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nature genetics 43(8), 761–767. [PubMed: 21743469]

Fan J, Fan Y, and Barut E (2014). Adaptive robust variable selection. The Annals of Statistics 42(1), 324–351. [PubMed: 25580039]

Fan J, Samworth R, and Wu Y (2009). Ultrahigh dimensional feature selection: beyond the linear model. Journal of Machine Learning Research 10, 2013–2038. [PubMed: 21603590]

Fang EX, Ning Y, and Liu H (2017). Testing and confidence intervals for high dimensional proportional hazards models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(5), 1415–1437.

Fei Z, Zhu J, Banerjee M, and Li Y (2019). Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach. Biometrics 75(2), 551–561. [PubMed: 30549000]

Fleming TR and Harrington DP (2011). Counting Processes and Survival Analysis, Volume 169. John Wiley & Sons.

Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, et al. (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS genetics 4(4), e1000054. [PubMed: 18437204]

Gulati S, Martinez P, Joshi T, Birkbak NJ, Santos CR, Rowan AJ, et al. (2014). Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. European urology 66(5), 936–948. [PubMed: 25047176]

Hanawa H, Hargrove PW, Kepes S, Srivastava DK, Nienhuis AW, and Persons DA (2004). Extended β-globin locus control region elements promote consistent therapeutic expression of a γ-globin lentiviral vector in murine β-thalassemia. Blood 104(8), 2281–2290. [PubMed: 15198957]

He X, Wang L, and Hong HG (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. The Annals of Statistics 41(1), 342–369.

Ho DSW, Schierding W, Wake M, Saffery R, and O'Sullivan J (2019). Machine learning SNP based prediction for precision medicine. Frontiers in Genetics 10, 267. [PubMed: 30972108]

Hong HG, Christiani DC, and Li Y (2019). Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. Precision clinical medicine 2(2), 90–99. [PubMed: 31355047]

Huang Y-T, Heist RS, Chirieac LR, Lin X, Skaug V, Zienolddiny S, et al. (2009). Genome-wide analysis of survival in early-stage non–small-cell lung cancer. Journal of clinical oncology 27(16), 2660–2667. [PubMed: 19414679]

Javanmard A and Montanari A (2014). Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15(1), 2869–2909.

Kelley MJ, Li S, and Harpole DH (2001). Genetic analysis of the β-tubulin gene, tubb, in non-small-cell lung cancer. Journal of the National Cancer Institute 93(24), 1886–1888. [PubMed: 11752014]

Koenker R and Bassett G Jr (1978). Regression quantiles. Econometrica: Journal of the Econometric Society 46(1), 33–50.

Kong S, Yu Z, Zhang X, and Cheng G (2018). High dimensional robust inference for cox regression models. arXiv preprint arXiv:1811.00535.

Korpanty GJ, Graham DM, Vincent MD, and Leighl NB (2014). Biomarkers that currently affect clinical practice in lung cancer: EGFR, ALK, MET, ROS-1, and KRAS. Frontiers in oncology 4, 204. [PubMed: 25157335]

Lee JD, Sun DL, Sun Y, and Taylor JE (2016). Exact post-selection inference, with application to the lasso. The Annals of Statistics 44(3), 907–927.

Lord RV, Brabender J, Gandara D, Alberola V, Camps C, Domine M, et al. (2002). Low ERCC1 expression correlates with prolonged survival after cisplatin plus gemcitabine chemotherapy in non-small cell lung cancer. Clinical Cancer Research 8(7), 2286–2291. [PubMed: 12114432]

McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature genetics 49(7), 1126–1132. [PubMed: 28604730]

Meinshausen N, Meier L, and Bühlmann P (2009). P-values for high-dimensional regression. Journal of the American Statistical Association 104(488), 1671–1681.

Moon C, Oh Y, and Roth JA (2003). Current status of gene therapy for lung cancer and head and neck cancer. Clinical cancer research 9(14), 5055–5067. [PubMed: 14613982]

Ning Y and Liu H (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. The Annals of Statistics 45(1), 158–195.

Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. (2004). Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304(5676), 1497–1500. [PubMed: 15118125]

Pâques F and Duchateau P (2007). Meganucleases and dna double-strand break-induced recombination: perspectives for gene therapy. Current gene therapy 7(1), 49–66. [PubMed: 17305528]

Peng L and Huang Y (2008). Survival analysis with quantile regression models. Journal of the American Statistical Association 103(482), 637–649.

Portnoy S (2003). Censored regression quantiles. Journal of the American Statistical Association 98(464), 1001–1012.

Powell JL (1986). Censored regression quantiles. Journal of econometrics 32(1), 143–155.

Risch A and Plass C (2008). Lung cancer epigenetics and genetics. International Journal of Cancer 123(1), 1–7. [PubMed: 18425819]

Rosell R, Molina MA, Costa C, Simonetti S, Gimenez-Capitan A, Bertran-Alamillo J, et al. (2011). Pretreatment EGFR T790M mutation and BRCA1 mRNA expression in erlotinib-treated advanced non–small-cell lung cancer patients with EGFR mutations. Clinical Cancer Research 17(5), 1160–1168. [PubMed: 21233402]

Sasaki H, Shimizu S, Endo K, Takada M, Kawahara M, Tanaka H, et al. (2006). EGFR and erbB2 mutation status in japanese lung cancer patients. International Journal of Cancer 118(1), 180–184. [PubMed: 16003726]

Shows JH, Lu W, and Zhang HH (2010). Sparse estimation and inference for censored median regression. Journal of Statistical Planning and Inference 140(7), 1903–1917. [PubMed: 20607110]

Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. (2007). Identification of the transforming eml4–alk fusion gene in non-small-cell lung cancer. Nature 448(7153), 561–566. [PubMed: 17625570]

Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, et al. (2012). RET, ROS1 and ALK fusions in lung cancer. Nature Medicine 18(3), 378–381.

Toyooka S, Tsuda T, and Gazdar AF (2003). The TP53 gene, tobacco exposure, and lung cancer. Human Mutation 21(3), 229–239. [PubMed: 12619108]

Van de Geer S, Bühlmann P, Ritov Y, and Dezeure R (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42(3), 1166–1202.

Wager S and Athey S (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523), 1228–1242.

Wager S, Hastie T, and Efron B (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. Journal of Machine Learning Research 15(1), 1625–1651. [PubMed: 25580094]

Wang L, Wu Y, and Li R (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association 107(497), 214–222. [PubMed: 23082036]

Wei L-J (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine 11(14–15), 1871–1879. [PubMed: 1480879]

Yamamoto H, Shigematsu H, Nomura M, Lockwood WW, Sato M, Okumura N, et al. (2008). Pik3ca mutations and copy number gains in human lung cancers. Cancer research 68(17), 6913–6921. [PubMed: 18757405]

Zhang C-H and Zhang SS (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1), 217–242.

Zhao P and Yu B (2006). On model selection consistency of lasso. Journal of Machine Learning Research 7(Nov), 2541–2563.

Zhao SD and Li Y (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis 105(1), 397–411. [PubMed: 22408278]

Zheng Q, Gallagher C, and Kulasekera K (2013). Adaptive penalized quantile regression for high dimensional data. Journal of Statistical Planning and Inference 143(6), 1029–1038.

Zheng Q, Peng L, and He X (2015). Globally adaptive quantile regression with ultra-high dimensional data. The Annals of Statistics 43(5), 2225–2258. [PubMed: 26604424]

Zheng Q, Peng L, and He X (2018). High dimensional censored quantile regression. The Annals of Statistics 46(1), 308–343. [PubMed: 30344355]

Zheng Z, Chen T, Li X, Haura E, Sharma A, and Bepler G (2007). DNA synthesis and repair genes RRM1 and ERCC1 in lung cancer. New England Journal of Medicine 356(8), 800–808. [PubMed: 17314339]

Zhu Q-G, Zhang S-M, Ding X-X, He B, and Zhang H-Q (2017). Driver genes in non-small cell lung cancer: Characteristics, detection methods, and targeted therapies. Oncotarget 8(34), 57680–57692. [PubMed: 28915704]

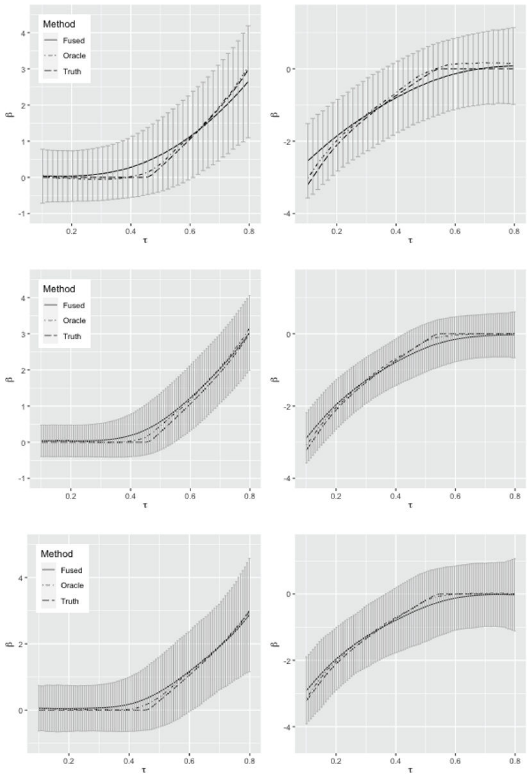

Figure 1:

Estimated heterogeneous effects and confidence intervals of Fused-HDCQR using Example 3: $\boldsymbol{\beta}_2^*(\cdot)$(left panel) and $\boldsymbol{\beta}_5^*(\cdot)$ (right panel). From the top to the bottom are the plots for $(n, p) = (300, 1000)$, $(700, 1000)$ and $(700, 2000)$, respectively.

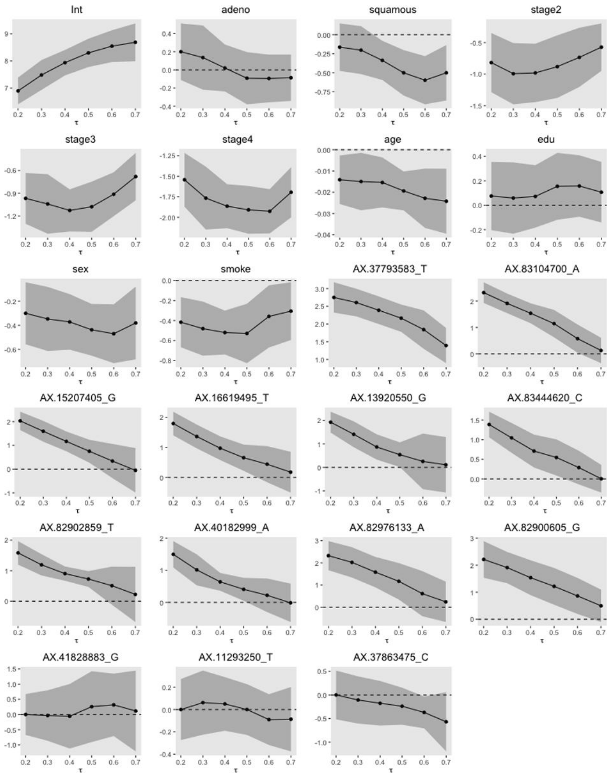

Figure 2:

Estimated quantile-specific coefficients of the predictors in Table 7.

Table 1:

Summary of variable selection results based on the simulated datasets.

				TP			FP		
	(n,p)	**CR**	**q**	**L-HDCQR**	**M09**	**F09**	**L-HDCQR**	**M09**	**F09**
Example 1	(300,1000)	0.25	3	2.67	2.12	1.64	7.95	0.00	0.19
	(700,1000)	0.25	3	2.98	2.78	2.27	13.08	0.01	0.34
Example 2	(300,1000)	0.22	4	3.60	3.58	2.22	12.45	0.00	0.22
	(700,1000)	0.23	4	3.99	3.99	3.54	11.29	0.00	0.64
Example 3	(300,1000)	0.20	5	3.82	3.63	1.91	10.00	0.00	0.17
	(700,1000)	0.20	5	4.81	4.77	4.35	11.73	0.01	0.54
	(700,2000)	0.19	5	4.78	4.76	4.17	16.34	0.00	0.47

Note: CR, average censoring rate; $q = |\mathcal{S}^*|$; TP, average true positives; FP, average false positives; M09, Meinshausen et al. (2009); F09, Fan et al. (2009); L-HDCQR, Zheng et al. (2018).

Table 2:

Results of Example 1 based on the simulated datasets.

	Bias					EmpSD	SE	Cov	Power	
	Oracle	Fused	Z18	F14	W12	Fused			Fused	M09
$n = 300$, $p = 1000$										
$\beta_{21} = 0.5$	0.02	0.02	−0.38	−0.50	−0.48	0.14	0.13	0.93	0.97	0.06
	0.02	0.01	−0.24	−0.49	−0.48	0.12	0.13	0.95	0.98	0.04
	0.01	0.01	−0.13	−0.50	−0.48	0.12	0.13	0.96	1.00	0.02
$\beta_{41} = 1$	−0.01	−0.01	−0.02	−0.91	−0.33	0.14	0.13	0.92	1.00	0.99
	−0.00	−0.00	−0.03	−0.68	−0.32	0.14	0.12	0.92	1.00	0.98
	0.02	0.01	−0.01	−0.70	−0.30	0.17	0.14	0.93	1.00	0.94
$\beta_{61} = 1.5$	−0.00	0.01	0.00	−0.92	−0.24	0.12	0.13	0.92	1.00	1.00
	0.00	0.01	0.01	−0.64	−0.25	0.11	0.13	0.97	1.00	1.00
	0.02	0.01	0.02	−0.70	−0.25	0.13	0.14	0.95	1.00	1.00
$n = 700$, $p = 1000$										
$\beta_{21} = 0.5$	−0.02	−0.01	−0.01	−0.47	−0.23	0.09	0.08	0.92	1.00	0.56
	−0.01	−0.01	−0.01	−0.39	−0.22	0.08	0.08	0.89	1.00	0.65
	−0.01	−0.01	−0.01	−0.40	−0.23	0.10	0.09	0.89	1.00	0.44
$\beta_{41} = 1$	0.00	0.00	0.04	−0.53	−0.17	0.09	0.08	0.91	1.00	1.00
	−0.00	0.00	0.03	−0.49	−0.19	0.09	0.08	0.90	1.00	1.00
	−0.01	−0.01	0.01	−0.53	−0.18	0.08	0.10	0.87	1.00	1.00
$\beta_{61} = 1.5$	0.01	0.01	0.06	−0.54	−0.21	0.10	0.09	0.93	1.00	1.00
	0.01	0.01	0.03	−0.62	−0.21	0.08	0.08	0.93	1.00	1.00
	−0.00	0.00	0.03	−0.71	−0.21	0.07	0.09	0.94	1.00	1.00

Note: Each β has three rows corresponding to τ = .25,.5,.75 from the top to bottom; EmpSD, empirical standard deviation; SE, average standard error; Cov, coverage probability; Oracle, Oracle estimator; Z18, Zheng et al. (2018).F14, Fan et al. (2014); W12, Wang et al. (2012); M09, Meinshausen et al. (2009).

Table 3:

Results of Example 2 based on the simulated datasets.

	Bias			EmpSD	SE	Cov	Freq	Power	
	Oracle	Fused	Z18		Fused			Fused	M09
				$n = 300$, $p = 1000$					
$\beta_4 = 1.5Q_\epsilon(\tau)$	0.01	0.13	0.29	0.32	0.31	0.88	0.73	0.82	0.16
	−0.05	−0.07	0.06	0.33	0.29	0.90		0.11	0.00
	0.01	−0.14	−0.05	0.31	0.34	0.82		0.62	0.10
$\beta_{21} = 1$	−0.01	−0.01	−0.01	0.14	0.13	0.90	0.69	1.00	0.88
	−0.03	−0.01	−0.05	0.12	0.12	0.91		1.00	0.92
	−0.01	−0.00	−0.02	0.14	0.13	0.92		1.00	0.84
$\beta_{41} = 1.5$	0.01	0.01	0.03	0.13	0.13	0.90	0.99	1.00	1.00
	−0.01	0.01	0.03	0.12	0.13	0.93		1.00	1.00
	−0.00	0.02	−0.02	0.13	0.14	0.93		1.00	1.00
$\beta_{61} = 2$	−0.03	−0.03	0.04	0.13	0.13	0.91	1.00	1.00	1.00
	−0.03	−0.02	0.03	0.11	0.13	0.92		1.00	1.00
	−0.01	−0.01	−0.00	0.12	0.15	0.95		1.00	1.00
				$n = 700$, $p = 1000$					
$\beta_4 = 1.5Q_\epsilon(\tau)$	0.03	0.08	0.19	0.19	0.21	0.92	0.89	0.99	0.61
	0.02	0.03	0.14	0.18	0.19	0.89		0.11	0.00
	0.04	−0.03	−0.01	0.21	0.23	0.92		0.97	0.56
$\beta_{21} = 1$	0.01	0.01	0.05	0.09	0.08	0.94	0.99	1.00	1.00
	0.01	0.01	0.01	0.08	0.08	0.87		1.00	1.00
	0.01	0.01	0.05	0.10	0.09	0.89		1.00	1.00
$\beta_{41} = 1.5$	−0.01	0.00	0.08	0.08	0.08	0.94	1.00	1.00	1.00
	−0.00	0.00	0.05	0.09	0.08	0.92		1.00	1.00
	0.00	0.01	0.04	0.09	0.09	0.95		1.00	1.00
$\beta_{61} = 2$	−0.01	−0.01	0.10	0.08	0.09	0.93	1.00	1.00	1.00
	−0.01	−0.01	0.06	0.08	0.09	0.91		1.00	1.00
	−0.00	−0.00	0.07	0.09	0.10	0.90		1.00	1.00

Note: See the footnote of Table 2; Freq, average selection frequency in B splits.

Table 4:

Results of Example 3 based on the simulated datasets.

	Bias			EmpSD	SE	Cov	Freq	Power	
	Oracle	Fused	Z18	Fused				Fused	M09
				$n = 300$, $p = 1000$					
$\beta_2 = \varphi_1(\tau)$	−0.05	0.06	0.59	0.34	0.36	0.94	0.71	0.06	0.00
	0.11	0.37	1.01	0.52	0.51	0.89		0.20	0.00
	0.04	−0.20	−0.05	0.80	0.72	0.89		0.87	0.06
$\beta_{11} = \varphi_{10}(\tau)$	0.08	0.14	0.27	0.65	0.50	0.90	0.67	0.77	0.36
	0.10	−0.20	−0.36	0.62	0.51	0.91		0.19	0.00
	0.16	0.06	−0.03	0.56	0.52	0.90		0.10	0.00
$\beta_{21} = 1.5$	0.03	0.03	0.04	0.25	0.23	0.95	0.65	1.00	0.77
$\beta_{41} = 2$	0.00	−0.00	0.02	0.23	0.25	0.93	0.93	1.00	0.99
$\beta_{61} = 2.5$	0.09	0.07	0.19	0.21	0.26	0.94	0.99	1.00	1.00
				$n = 700$, $p = 1000$					
$\beta_2 = \varphi_1(\tau)$	0.02	0.04	0.27	0.21	0.23	0.94	0.96	0.06	0.00
	0.17	0.30	0.79	0.37	0.40	0.88		0.27	0.01
	0.15	0.08	0.35	0.51	0.51	0.90		1.00	0.77
$\beta_{11} = \varphi_{10}(\tau)$	0.07	0.09	0.18	0.33	0.33	0.91	0.92	0.99	0.92
	−0.01	−0.19	−0.23	0.35	0.34	0.85		0.21	0.00
	−0.00	−0.04	−0.08	0.37	0.31	0.94		0.06	0.00
$\beta_{21} = 1.5$	−0.00	0.00	0.04	0.16	0.17	0.97	0.98	1.00	1.00
$\beta_{41} = 2$	−0.03	−0.02	−0.01	0.15	0.18	0.95	1.00	1.00	1.00
$\beta_{61} = 2.5$	0.00	0.00	0.07	0.18	0.18	0.94	1.00	1.00	1.00
				$n = 700$, $p = 2000$					
$\beta_2 = \varphi_1(\tau)$	0.05	0.11	0.13	0.32	0.32	0.93	0.93	0.07	0.00
	0.09	0.34	0.87	0.46	0.44	0.91		0.09	0.02
	0.25	0.36	1.77	0.53	0.46	0.87		0.74	0.58
$\beta_{11} = \varphi_{10}(\tau)$	0.13	0.25	0.73	0.45	0.35	0.84	0.90	1.00	0.83
	0.09	−0.02	0.56	0.41	0.36	0.89		0.76	0.01
	−0.04	−0.30	−0.13	0.36	0.34	0.85		0.15	0.00
$\beta_{21} = 1.5$	0.01	0.01	0.03	0.18	0.21	0.98	0.98	1.00	1.00
$\beta_{41} = 2$	0.01	0.03	−0.07	0.22	0.20	0.91	0.99	1.00	0.98
$\beta_{61} = 2.5$	−0.02	−0.01	−0.05	0.25	0.20	0.94	1.00	1.00	0.98

Note: See the footnote of Tables 2 and 3; For β_2 and β_{11}, the numbers are shown at $\tau = .25,.5,.75$ from the top to the bottom and, for the other β's, at $\tau = 0.5$.

Table 5:

Comparisons of average computing time (in seconds) when performing Example 1.

	Fused	Z18	W12	F14	M09
(n,p) = (300,1000)	888	853	509	390	170
(n,p) = (700,1000)	3,108	1,812	2,230	1,231	440

Note: see the footnote of Table 2.

Table 6:

Patients' characteristics in the BLCSC samples.

(*n* = 674)		
		Mean (SD)
Age		60 (10.8)
		Count (%)
Female		259 (38)
Education level	≤ High school	264 (39)
	> High school	410 (61)
Smoking	Non-active	418 (62)
	Active	256 (38)
Cancer type	Adenocarcinoma	283 (42)
	Squamous cell	110 (16)
	Other	281 (42)
Cancer stage	1	283 (42)
	2	110 (16)
	3	256 (38)
	4	25 (4)

Table 7:

Analysis of the BLCSC data with Fused-HDCQR. The SNPs are sorted by their *p*-values at $\tau = 0{:}2$, corresponding to the high risk groups. Results for the top 10 and the bottom 3 are presented.

	Estimator	SE	*p*-value	Estimator	SE	*p*-value	Estimator	SE	*p*-value
τ		0.2			0.3			0.4	
Int	6.90	0.25	1.4E–165	7.48	0.28	4.3E–157	7.94	0.24	3.2E–241
Adeno	0.20	0.16	2.1E–01	0.14	0.18	4.5E–01	0.02	0.13	8.7E–01
Squamous	−0.16	0.16	3.0E–01	−0.20	0.16	2.1E–01	−0.34	0.13	1.0E–02
Stage2	−0.82	0.24	6.3E–04	−0.99	0.25	6.0E–05	−0.98	0.24	3.2E–05
Stage3	−0.97	0.17	1.6E–08	−1.04	0.20	2.0E–07	−1.13	0.14	2.0E–15
Stage4	−1.54	0.17	3.0E–20	−1.77	0.20	1.7E–19	−1.86	0.14	2.2E–42
Age	−0.01	0.01	1.5E–02	−0.01	0.01	3.0E–02	−0.02	0.01	1.0E–02
Edu	0.08	0.14	6.0E–01	0.06	0.15	6.9E–01	0.07	0.13	5.8E–01
Female	−0.30	0.13	2.2E–02	−0.35	0.14	1.0E–02	−0.37	0.12	1.6E–03
Smoke	−0.42	0.13	1.1E–03	−0.48	0.14	5.0E–04	−0.52	0.11	3.4E–06
AX.37793583 T	2.75	0.22	3.0E–36	2.61	0.20	4.6E–39	2.39	0.20	3.7E–33
AX.83104700 A	2.32	0.20	4.0E–31	1.91	0.19	6.3E–24	1.54	0.19	1.5E–15
AX.15207405 G	2.03	0.20	1.0E–24	1.59	0.22	9.8E–13	1.17	0.21	3.7E–08
AX.16619495 T	1.79	0.20	3.3E–19	1.36	0.20	1.3E–11	0.97	0.20	1.2E–06
AX.13920550 G	1.93	0.23	2.5E–17	1.41	0.28	5.3E–07	0.87	0.27	1.6E–03
AX.83444620 C	1.39	0.17	7.4E–17	1.05	0.19	6.6E–08	0.71	0.21	8.8E–04
AX.82902859 T	1.58	0.20	8.7E–16	1.19	0.18	2.0E–11	0.90	0.12	3.4E–14
AX.40182999 A	1.50	0.21	9.6E–13	1.01	0.25	3.9E–05	0.64	0.14	6.5E–06
AX.82976133 A	2.32	0.33	3.8E–12	2.02	0.35	6.7E–09	1.58	0.35	6.1E–06
AX.82900605 G	2.21	0.35	1.6E–10	1.91	0.29	9.1E–11	1.54	0.33	2.9E–06
...									
AX.41828883 G	1.4E–03	0.34	1.00	−3.2E–02	0.42	0.94	−5.7E–02	0.54	0.92
AX.11293250 T	−3.6E–04	0.14	1.00	6.2E–02	0.15	0.67	5.0E–02	0.12	0.68
AX.37863475 C	−3.1E–04	0.26	1.00	−1.1E–01	0.25	0.68	−1.8E–01	0.24	0.46
Int	8.30	0.27	4.8E-214	8.55	0.30	4.9E-180	8.69	0.35	2.8E-132
Adeno	−0.09	0.15	5.3E–01	−0.09	0.13	4.8E–01	−0.09	0.13	5.1E–01

	Estimator	SE	*p*-value	Estimator	SE	*p*-value	Estimator	SE	*p*-value
τ		**0.2**			**0.3**			**0.4**	
Squamous	−0.50	0.15	1.0E–03	−0.60	0.16	2.1E–04	−0.50	0.19	7.1E–03
Stage2	−0.88	0.25	5.0E–04	−0.73	0.24	2.1E–03	−0.57	0.19	2.8E–03
Stage3	−1.08	0.17	1.7E–10	−0.91	0.15	6.4E–10	−0.68	0.16	2.0E–05
Stage4	−1.91	0.15	7.0E–38	−1.93	0.14	1.7E–44	−1.69	0.16	2.1E–27
Age	−0.02	0.00	3.3E–05	−0.02	0.01	1.3E–03	−0.02	0.01	1.9E–03
Edu	0.15	0.14	2.7E–01	0.16	0.13	2.2E–01	0.11	0.13	4.0E–01
Female	−0.44	0.11	6.4E–05	−0.47	0.12	1.6E–04	−0.38	0.15	1.3E–02
Smoke	−0.53	0.15	4.9E–04	−0.36	0.16	2.4E–02	−0.31	0.15	3.8E–02
AX.37793583 T	2.16	0.20	4.1E–28	1.84	0.28	2.8E–11	1.39	0.25	4.2E–08
AX.83104700 A	1.15	0.27	1.6E–05	0.58	0.27	3.5E–02	0.13	0.25	6.0E–01
AX.15207405 G	0.75	0.25	2.3E–03	0.34	0.37	3.5E–01	−0.05	0.48	9.2E–01
AX.16619495 T	0.66	0.22	3.1E–03	0.44	0.31	1.5E–01	0.18	0.35	6.1E–01
AX.13920550 G	0.54	0.27	4.3E–02	0.26	0.60	6.7E–01	0.11	0.60	8.6E–01
AX.83444620 C	0.55	0.23	2.0E–02	0.29	0.22	1.8E–01	0.01	0.18	9.7E–01
AX.82902859 T	0.73	0.13	4.2E–08	0.51	0.32	1.1E–01	0.22	0.46	6.3E–01
AX.40182999 A	0.41	0.18	2.6E–02	0.22	0.27	4.1E–01	−0.01	0.30	9.6E–01
AX.82976133 A	1.17	0.42	5.4E–03	0.61	0.52	2.4E–01	0.24	0.46	6.0E–01
AX.82900605 G	1.22	0.35	4.5E–04	0.86	0.34	1.1E–02	0.50	0.31	1.0E–01
...									
AX.41828883 G	0.26	0.60	0.66	0.32	0.52	0.54	0.12	0.68	0.86
AX.11293250 T	−0.00	0.12	1.00	−0.09	0.12	0.44	−0.09	0.15	0.56
AX.37863475 C	−0.24	0.20	0.23	−0.37	0.17	0.03	−0.57	0.32	0.08