Assessing Quantile Prediction With Censored Quantile Regression Models

Ruosha Li[1] and **Limin Peng**[2]

[1]Department of Biostatistics, The University of Texas School of Public Health, Houston, TX

[2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

Summary

An important goal of censored quantile regression is to provide reliable predictions of survival quantiles, which are often reported in practice to offer robust and comprehensive biomedical summaries. However, formal methods for evaluating and comparing *working* quantile regression models in terms of their performance in predicting survival quantiles have been lacking, especially when the working models are subject to model mis-specification. In this paper, we proposes a sensible and rigorous framework to fill in this gap. We introduce and justify a predictive performance measure defined based on the check loss function. We derive estimators of the proposed predictive performance measure and study their distributional properties and the corresponding inference procedures. More importantly, we develop model comparison procedures that enable thorough evaluations of model predictive performance among nested or non-nested models. Our proposals properly accommodate random censoring to the survival outcome and the realistic complication of model mis-specification, and thus are generally applicable. Extensive simulations and a real data example demonstrate satisfactory performances of the proposed methods in real life settings.

Keywords

1. INTRODUCTION

Quantile regression (Koenker and Bassett, 1978) has emerged as a useful approach to analyzing survival data. It permits comprehensive explorations of covariate effects across different segments of an event time distribution. Moreover, it serves as a major tool for predicting survival quantiles. Quantiles of a survival time have been frequently reported in biomedical studies; they have many appealing features including straightforward interpretations and invariance to monotone transformations. Moreover, while the mean of a survival time is often not identifiable when censoring presents, the quantiles of a survival time can still be estimated at a range of quantile levels. Accurate predictions of survival

---

quantiles can provide valuable biomedical insights, and constitute a fundamental goal of applying quantile regression to survival data.

Most of research efforts on quantile regression with survival data have been oriented to developing estimation and inference procedures for models assumed to perfectly conform to the underlying data mechanism (Ying et al. 1995; Portnoy 2003; Zhou 2006; Peng and Huang 2008; Wang and Wang 2009; Huang 2010; among others). However, very limited attention was paid to the evaluation of quantile regression models as well as the comparison among a set of candidate models.

Existing methods for assessing quantile regression models are generally focused on models' goodness-of-fit or lack-of-fit. For example, Koenker and Machado (1999) proposed a goodness-of-fit criterion based on the check loss function $\rho_\tau(u) = u\{\tau - I(u \quad 0)\}$, which is analogous to the R^2 statistic in classical least squares regression. He and Zhu (2003) proposed an omnibus lack-of-fit test for linear or nonlinear quantile regression models. Wang (2008) studied a nonparametric test for checking the lack-of-fit of censored quantile regression. All these methods assess models based on their fit to existing data.

The perspective we take in this paper is to evaluate or compare quantile regression models based on their capacity in predicting quantiles of *future* outcomes. Further, we do not assume that we know the form of the true model that explains the underlying relationship between covariates and response. This means, we view any adopted quantile regression model as a *working* model. Similar model evaluation frameworks have been studied in other settings, where the quantities to predict do not relate to quantiles. For example, Tian et al. (2007) used absolute prediction error to assess the predictive performance of working linear transformation models for an uncensored continuous response. With survival data, Uno et al. (2007) developed a framework for evaluating a model's predictive performance for t-year survival status. Tian et al. (2014) considered evaluating the prediction performance for the restricted mean event time (RMET), where the RMET can be estimated either semi-parametrically or non-parametrically. Such works shed useful insight for our proposals in this paper, which target the prediction of quantiles.

To assess the prediction of quantiles, an intuitive idea is to evaluate the squared or absolute difference between true and predicted conditional quantiles. However true conditional quantiles are not directly observable. This adds subtlety to the assessment of the predictive performance of quantile regression models. Moreover, to evaluate quantile prediction with survival data, censoring is an important feature that needs to be appropriately handled, like in the model estimation setting. Noh et al. (2013) made a precursor effort for the uncensored setting by studying the properties of Koenker and Machado (1999)'s model adequacy measures in the presence of model mis-specification. Their work provided an approach to comparing two nested linear quantile models regardless of whether the linear specification of the conditional quantile functions is correct or not. It is tempting to develop a formal framework for assessing quantile prediction that is suitable for censored responses and allows for more general types of quantile predictions and model specification.

$\xi_\tau^0 \equiv Q_{T^0}(\tau|\overline{\mathbf{Z}}^0)$ denote the true conditional quantile of T^0 of the new subject. Following the arguments in Koenker (2005) (Section 2.9) and Angrist et al. (2006) and temporarily suppressing $\hat{\xi}_\tau(\mathbf{Z}^0)$ to $\hat{\xi}_\tau$, we can show that

$$E\{\rho_\tau(T^0 - \hat{\xi}_\tau) - \rho_\tau(T^0 - \xi_\tau^0)|\overline{\mathbf{Z}}^0, \mathscr{D}_m\} = \int_0^{\hat{\xi}_\tau - \xi_\tau^0}\{F(\xi_\tau^0 + s|\overline{\mathbf{Z}}^0) - \tau\}ds, \quad (1)$$

where $F(\cdot|\overline{\mathbf{Z}}^0)$ is the conditional distribution function of T^0 given $\overline{\mathbf{Z}}^0$. Let $d_{\overline{\mathbf{z}}^0}(\xi_\tau) = \int_0^{\xi_\tau - \xi_\tau^0}\{F(\xi_\tau^0 + s|\overline{\mathbf{Z}}^0) - \tau\}ds$, which is always non-negative. Let $f(\cdot|\overline{\mathbf{Z}}^0)$ be the conditional density function of T^0 given $\overline{\mathbf{Z}}^0$. When $f(\cdot|\overline{\mathbf{Z}}^0)$ is bounded, $d_{\overline{\mathbf{z}}^0}(\xi_\tau) = \frac{1}{2}f(\xi_\tau^0|\overline{\mathbf{Z}}^0)(\xi_\tau - \xi_\tau^0)^2 + o\{(\xi_\tau - \xi_\tau^0)^2\}$. Let $E_{\overline{\mathbf{Z}}^0}$ denote expectation taken over $\overline{\mathbf{Z}}^0$, we see that

$$E_{\overline{\mathbf{z}}^0}\{\rho_\tau(T^0 - \hat{\xi}_\tau) - \rho_\tau(T^0 - \xi_\tau^0)|\mathscr{D}_m\} = E_{\overline{\mathbf{z}}^0}\{d_{\overline{\mathbf{z}}^0}(\hat{\xi}_\tau)|\mathscr{D}_m\},$$

and approximates $E_{\overline{\mathbf{z}}^0}\{\frac{1}{2}f(\xi_\tau^0|\overline{\mathbf{Z}}^0)(\hat{\xi}_\tau - \xi_\tau^0)^2|\mathscr{D}_m\}$, which may be viewed as a weighted squared difference between the predicted quantile $\hat{\xi}_\tau$ and the true quantile ξ_τ^0 with weight $\frac{1}{2}f(\xi_\tau^0|\overline{\mathbf{Z}}^0)$. Note that the true quantile ξ_τ^0 and so $E\{\rho_\tau(T^0 - \xi_\tau^0)\}$ are generally unknown in practice but are fixed. Hence assessing $E\{\rho_\tau(T^0 - \hat{\xi}_\tau)|\mathscr{D}_m\}$ allows us to evaluate how well $\hat{\xi}_\tau$ predicts ξ_τ^0 *relatively* to predictors from other models.

The ultimate goal here is to assess the capacity of a working model in predicting the true quantiles. To this end, we propose to take a further step by considering $\lim_{m\to\infty} E\rho_\tau\{T^0 - \hat{\xi}_\tau(\mathbf{Z}^0)|\mathscr{D}_m\}$. For a working model to serve as a reliable prediction tool, it is desirable that the estimated parameters in the model converge in probability to deterministic values when m increases, even under potential model mis-specification. As a result, the predictor $\hat{\xi}_\tau(\cdot)$ will also converge in probability to a deterministic function $\tilde{\xi}(\cdot)$ uniformly in $\mathbf{Z}^0$, and thus $\lim_{m\to\infty} E\rho_\tau\{T^0 - \hat{\xi}_\tau(\mathbf{Z}^0)|\mathscr{D}_m\} = E\rho_\tau\{T^0 - \tilde{\xi}_\tau(\mathbf{Z}^0)\}$. This quantity essentially captures the predictive capacity of a working model at the τ_{th} quantile.

2.2 Quantile prediction loss with censored outcomes

We now consider right-censored survival data. Let T represent the log survival time and C be the log censoring time. In the presence of censoring, the identifiability of $E\rho_\tau\{T^0 - \xi_\tau(\mathbf{Z}^0)\}$ is often of concern. This is particularly true when the upper bound of the censoring support is less than that of T's support in the observed data. It is natural to consider a modified check loss measure for censored data as $E[\rho_\tau\{T^{0u} - \xi_\tau(\mathbf{Z}^0)\}]$, where $T^{0u} = T^0 \wedge u$ and u is a prespecified *deterministic* constant that is less than the upper bound of C's support. For example, in many clinical studies, u can be chosen to be slightly smaller than the planned follow-up time in the study protocol. Similar truncation techniques have been adopted in many other survival settings, for example, studying restricted mean (Zucker, 1998; Chen and

Tsiatis, 2001; Goldberg and Kosorok, 2012) and evaluating predictive performance (Tian et al., 2014; Lawless and Yuan, 2010). When $\tau = 0.5$, $L(\tau, \boldsymbol{\beta}) = 0.5|T^{0u} - \xi_\tau(\boldsymbol{Z}^0)|$ and is quite similar to the absolute prediction loss in Tian et al. (2014) for the RMET.

Write $Y = T \wedge C$ and $\delta = I(T \quad C)$. Suppose that $\hat{\xi}_\tau(\cdot)$ is an estimated quantile predictor for $Q_T(\tau|\bar{\boldsymbol{Z}})$, derived under a working model using dataset $\mathscr{D}_m \equiv (Y_i, \delta_i, \bar{\boldsymbol{Z}}_i)_{i=1}^m$. In this work, we consider linear quantile regression models as the working models to obtain $\hat{\xi}_\tau(\cdot)$. Standard survival models, including the Cox proportional hazards model, can also serve as reasonable working models. For example, under a Cox proportional hazards model, $\hat{\xi}_\tau(\boldsymbol{Z})$ can be defined as $\log [\inf \{t : \exp\{-\hat{H}_0(t) \exp(\boldsymbol{Z}\hat{\boldsymbol{b}})\} \quad 1 - \tau\}]$, where $\hat{H}_0(t)$ and $\hat{\boldsymbol{b}}$ are the estimated baseline cumulative hazard function and the Cox regression coefficient respectively. Methods we propose below can be easily adapted to situations with such choices of $\hat{\xi}_\tau(\cdot)$.

A common linear quantile regression model takes the form,

$$Q_T(\tau|\bar{\boldsymbol{Z}}) = \boldsymbol{Z}\boldsymbol{\beta}_0(\tau), \tau \in [\tau_L, \tau_U], \quad (2)$$

where $\boldsymbol{Z}$ is a subvector of $\bar{\boldsymbol{Z}}$ or $\bar{\boldsymbol{Z}}$ itself, including 1 as the first component, and $0 < \tau_L \quad \tau_U < 1$. Using the dataset $\mathscr{D}_m$ well representing the study population of interest, one may implement an existing quantile regression procedure to obtain an estimator of $\boldsymbol{\beta}_0(\tau)$, denoted by $\hat{\boldsymbol{\beta}}(\tau)$. For a *new* subject from the same population with $\bar{\boldsymbol{Z}} = \bar{\boldsymbol{Z}}^0$, based on model (2), a prediction of the τ_{th} quantile of the potential log survival time T^0 is given by $\hat{\xi}_\tau(\boldsymbol{Z}^0) = \boldsymbol{Z}^0\hat{\boldsymbol{\beta}}(\tau)$. In this work, we shall focus on the *log* survival times, but the methods can also handle the original survival times with minor modifications.

Under model (2), the quantile prediction loss becomes $L(\tau, \hat{\boldsymbol{\beta}})$, where

$$L(\tau, \boldsymbol{\beta}) \equiv E\left[\rho_\tau\{T^{0u} - \boldsymbol{Z}^0\beta(\tau)\}\right], \quad (3)$$

is a continuous function of $\boldsymbol{\beta}(\tau)$ at a fixed τ. When sample size increases, a desirable $\hat{\boldsymbol{\beta}}(\tau)$ is expected to converge in probability to a deterministic function of τ, denoted by $\tilde{\boldsymbol{\beta}}(\tau)$. Consequently, $L(\tau, \hat{\boldsymbol{\beta}})$ will also converge in probability to its limiting value $L(\tau) \equiv L(\tau, \tilde{\boldsymbol{\beta}})$, which may be more precisely expressed as $L(\tau, \tilde{\boldsymbol{\beta}}) = E[\rho_\tau\{T^{0u} - \boldsymbol{Z}^0\tilde{\boldsymbol{\beta}}(\tau)\}]$. In addition, we hope that the the selected $\hat{\boldsymbol{\beta}}(\tau)$ can lead to an $L(\tau, \tilde{\boldsymbol{\beta}})$ with meaningful interpretations. That is, the resulting $L(\tau, \tilde{\boldsymbol{\beta}})$ can capture the capacity of the working model in predicting conditional quantiles. As such, $L(\tau, \tilde{\boldsymbol{\beta}})$ provides a sensible approach to evaluating and comparing models. Based on the arguments in Section 2.1 on $E\{\rho_\tau(T^0 - \hat{\xi}_\tau) - \rho_\tau(T^0 - \xi_\tau^0)\}$, one should interpret $L(\tau, \tilde{\boldsymbol{\beta}})$ in a relative scale, for example, through comparisons between different working models.

In this work we particularly investigate the case with $\hat{\boldsymbol{\beta}}(\tau)$ chosen as Zhou (2006)'s estimator of $\boldsymbol{\beta}_0(\tau)$ for a working model of form (2), which minimizes

$$\hat{L}_m(\tau, \boldsymbol{b}) = m^{-1} \sum_{i=1}^{m} \frac{\Delta_i}{\hat{G}(Y_i^u | \boldsymbol{Z}_i)} \rho_\tau \{Y_i^u - \boldsymbol{Z}_i^T \boldsymbol{b}\}, \tag{4}$$

where $Y_i^u = T_i^u \wedge C_i = Y_i \wedge u$, $\Delta_i = I(T_i^u \le C_i) = I(Y_i > u) + I(Y_i \le u, \delta_i = 1)$, and $\hat{G}(t|\boldsymbol{Z})$ is a consistent estimator for $\Pr(C > t|\boldsymbol{Z})$. The validity of (4) requires $C \perp T$ given $\boldsymbol{Z}$, where $\perp$ stands for statistical independence. In addition, it only relies on the conditional quantile modeling at the single quantile level τ, not all quantile levels in $[\tau_L, \tau_U]$. When it is reasonable to assume the independence between C and $\boldsymbol{Z}$, $\hat{G}(\cdot|\boldsymbol{Z})$ can reduce to $\hat{G}(\cdot)$, the Kaplan-Meier estimator of the unconditional survival function of C. When C and $\boldsymbol{Z}$ are dependent, one may obtain $\hat{G}(\cdot|\boldsymbol{Z})$ after imposing a proper regression model for C given $\boldsymbol{Z}$, such as the Cox proportional hazards model. It is worth noting that the validity of the proposed estimation procedure requires a correct specification of the assumed regression model for C. Adopting an inadequate regression model for C may lead to biased $\hat{G}(\cdot|\boldsymbol{Z})$ and consequently biased estimation of $L(\tau, \tilde{\boldsymbol{\beta}})$. For presentation simplicity, we assume that $C \perp (T, \boldsymbol{Z})$ in the sequel.

As we discuss in Supplemental Material A.1, this $\hat{\boldsymbol{\beta}}(\tau)$ can be shown to converge in probability to a deterministic value $\tilde{\boldsymbol{\beta}}(\tau)$ even under model mis-specification, where $\tilde{\boldsymbol{\beta}}(\tau)$ coincides with the minimizer of $E\{\rho_\tau(T - \boldsymbol{Z}\boldsymbol{b})\}$ regarding $\boldsymbol{b}$. Thus, $\boldsymbol{Z}\tilde{\boldsymbol{\beta}}(\tau)$ can be viewed as the theoretically best predictor under the working model (2). In the rest of the paper, we focus on the utility of the $L(\tau, \tilde{\boldsymbol{\beta}})$ discussed above as a predictive performance measure.

2.3 Summary measure of prediction performance

Let $L(\tau)$ be the shorthand notation for $L(\tau, \tilde{\boldsymbol{\beta}})$, which captures the prediction loss locally at the τ_{th} quantile. If we have multiple working models, a comparison of the resulting $L(\tau)$'s can reveal the relative prediction loss of these models at the τ_{th} quantile. One caveat with $L(\tau)$ is that its magnitude generally depends on τ, making it difficult to compare $L(\tau)$'s over different τ's. To address this limitation, we propose *coefficient of determinant* measure, $R^1(\tau)$, for censored data. The measure is analogous to the R^2 coefficient for least squares regression, and has been studied for uncensored data (Koenker and Machado, 1999; Noh et al., 2013). With censored data, we propose to use

$$R^1(\tau, \tilde{\boldsymbol{\beta}}) = 1 - \frac{E\rho_\tau(T^{0u} - \boldsymbol{Z}^0 \tilde{\boldsymbol{\beta}})}{E\rho_\tau(T^{0u} - \zeta_\tau)} = \frac{L_0(\tau) - L(\tau, \tilde{\boldsymbol{\beta}})}{L_0(\tau)}, \tag{5}$$

where ζ_τ denotes the true unconditional τ_{th} quantile of T^0, and $L_0(\tau) \equiv E\rho_\tau(T^{0u} - \zeta_\tau)$. It is worth noting that this measure is different from existing goodness-of-fit measures, because T^0 and $\boldsymbol{Z}^0$ denote data from a new subject. Unlike $L(\tau)$, the scale of $R^1(\tau) \equiv R^1(\tau, \tilde{\boldsymbol{\beta}})$ does not vary across τ and $R^1(\tau)$ always lies between 0 and 1. An illustrative example of $L(\tau)$ and $R^1(\tau)$ is provided in Figure 1 of Section 5. Note that $R^1(\tau)$ quantifies the extent to which model (2) improves quantile prediction at the τ_{th} quantile, when compared to a degenerated model with only the intercept term. Therefore, $R^1(\tau)$ can serve as a sensible measure of the

performance in predicting the τ-th quantile. We also propose a summary measure as $R^1 = (\tau_U - \tau_L)^{-1} \int_{\tau_L}^{\tau_U} R^1(\tau) d\tau$, which represents the overall prediction performances of a model across $\tau \in [\tau_L, \tau_U]$.

3. ESTIMATION AND INFERENCE

3.1 The proposed estimators of $L(\tau, \tilde{\beta})$ and $R^1(\tau, \tilde{\beta})$

In this section, we study the estimation of $L(\tau, \tilde{\beta})$ and $R^1(\tau, \tilde{\beta})$, which are justified as appropriate predictive performance measures in Section 2. We first consider a plug-in type estimator for $L(\tau) \equiv L(\tau, \tilde{\beta})$. Suppose that the observed data consist of n i.i.d. observations $(Y_i, \delta_i, \overline{Z}_i)_{i=1}^n$, which contain the data $\mathscr{D}_m$ used to derive $\hat{\beta}$. The plug-in estimator is

$$\hat{L}_n(\tau, \hat{\beta}) = n^{-1} \sum_{i=1}^m \frac{\Delta_i}{\hat{G}(Y_i^u | Z_i)} \rho_\tau \{Y_i^u - Z_i \hat{\beta}(\tau)\}. \tag{6}$$

where $\tilde{\beta}(\tau)$ is the estimator in Zhou (2006) using all observed data. The remarks about the censoring mechanism and $\hat{G}(\cdot)$ in Section 2.2 also apply to the $\hat{G}(\cdot)$ here. In the following, we adopt the Kaplan-Meier estimator $\hat{G}(\cdot)$ under the $C \perp (T, Z)$ assumption for the purpose of illustration. The plug-in estimator for $R^1(\tau)$ is $\hat{R}_n^1(\tau, \hat{\beta}) = 1 - \hat{L}_n(\tau, \hat{\beta}) / \hat{L}_n(\tau, \hat{\zeta}_\tau)$, where $\hat{\zeta}_\tau$ represents Zhou (2006)'s estimator in an intercept-only model.

Assume that $(T^0, \overline{Z}_0)$ in the definition of $L(\tau, \tilde{\beta})$ comes from the same population. In Supplementary Materials A.1, we use empirical process techniques to show that $\hat{L}_n(\tau, \hat{\beta})$ and $\hat{R}_n^1(\tau, \hat{\beta})$ are uniformly consistent to $L(\tau)$ and $R^1(\tau)$ respectively. The simpler notation $L(\tau)$ and $R^1(\tau)$ are adopted for $L(\tau, \tilde{\beta})$ and $R^1(\tau, \tilde{\beta})$ unless confusions occur. The results are formally stated in Theorem 1, where $\|x\|$ denotes the Euclidean norm of a vector x.

Theorem 1—*Under regularity conditions* **C1–C4** *in* Supplementary Materials A, *we have* $\sup_{\tau \in [\tau_L, \tau_U]} \|\hat{\beta}(\tau) - \tilde{\beta}(\tau)\| \xrightarrow{p} 0$, *where* $\tilde{\beta}(\tau)$ *is the solution to* $E[Z' \{\Pr(T \ Zb | Z) - \tau\}] = 0$. *Furthermore*,

$$\sup_{\tau \in [\tau_L, \tau_U]} |\hat{L}_n(\tau, \hat{\beta}) - L(\tau)| \xrightarrow{p} 0; \qquad \sup_{\tau \in [\tau_L, \tau_U]} |\hat{R}_n^1(\tau, \hat{\beta}) - R^1(\tau)| \xrightarrow{p} 0. \tag{7}$$

To estimate $L(\tau, \tilde{\beta})$, the plug-in estimator $\hat{L}_n(\tau, \hat{\beta})$ uses the entire observed dataset twice; the first use is to use it as $\mathscr{D}_m$ for estimating $\tilde{\beta}(\tau)$ via (4) and the second use is for approximating the check loss function in (6). The downward bias caused by using the same data twice can be shown to be of order $O_p(n^{-1})$, which is not of concern with a reasonably large n such as $n = 400$. With a small n, a bias correction procedure can be implemented. More details and justifications for the bias correction is provided at the end of Section 4.1.

Alternatively, we consider a cross validation (CV) type estimator, where different subsets of the data are utilized for estimating $\tilde{\boldsymbol{\beta}}(\tau)$ and $L(\tau, \tilde{\boldsymbol{\beta}})$. Specifically, one may split the observed data randomly into K disjoint subsets, with approximately equal sample sizes close to n/K. Here, K is a small fixed integer that does not grow with n. In practice, we suggest using $K = 10$ or $K = 5$. Let $V_i \in \{1, 2, \ldots, K\}$ denote the subset membership of the ith observation. For each subset k, let $\hat{\boldsymbol{\beta}}_{(-k)}(\tau)$ represent the estimator of $\boldsymbol{\beta}_0(\tau)$ based on all observations except those in the k_{th} subset. We define

$$\hat{L}^k\{\tau, \hat{\boldsymbol{\beta}}_{(-k)}\} = \frac{K}{n}\sum_{i=1}^{n}\frac{\Delta_i I(V_i=k)}{\hat{G}(Y_i^u)}\rho_\tau\{Y_i^u - \boldsymbol{Z}_i\hat{\boldsymbol{\beta}}_{(-k)}(\tau)\}, k=1, 2, \ldots, K.$$

This loss function averages across the check loss among subjects in the kth subset, where the predicted quantile is obtained via $\hat{\boldsymbol{\beta}}_{(-k)}$. Finally, the proposed the CV-type estimator of $L(\tau, \tilde{\boldsymbol{\beta}})$ can be calculated as $\hat{L}_{CV}(\tau, \hat{\boldsymbol{\beta}}) = K^{-1}\sum_{k=1}^{K}\hat{L}^k\{\tau, \hat{\boldsymbol{\beta}}_{(-k)}\}$, and a corresponding estimator for $R^1(\tau)$ is $\hat{R}^1_{CV}(\tau, \hat{\boldsymbol{\beta}}) = 1 - \hat{L}_{CV}(\tau, \hat{\boldsymbol{\beta}})/\hat{L}_{CV}(\tau, \hat{\zeta}_\tau)$. Similar to the plug-in estimators, $\hat{L}_{CV}(\tau, \hat{\boldsymbol{\beta}})$ and $\hat{R}^1_{CV}(\tau, \hat{\boldsymbol{\beta}})$ provides uniformly consistent estimations for $L(\tau)$ and $R^1(\tau)$. The proof is provided in Supplementary Material A.3.

3.2 Distributional properties of the proposed estimators

It is also important to examine the distributional properties of the estimators, so as to enable statistical inferences on the predictive capacity measures. We derived the following theorem:

Theorem 2—*Under conditions* **C1–C4** *in* Supplementary Materials A, *we have*

$$\sqrt{n}\{\hat{L}_n(\tau, \hat{\boldsymbol{\beta}}) - L(\tau)\} = n^{-1/2}\sum_{i=1}^{n}\pi_i(\tau, \tilde{\boldsymbol{\beta}}) + o_p^\tau(1),$$

(8)

where $o_p^\tau(1)$ *stands for a function of* τ *converging to* 0 *in probability uniformly in* $\tau \in [\tau_L, \tau_U]$. *Furthermore,* $\sqrt{n}\{\hat{L}_n(\tau, \hat{\boldsymbol{\beta}}) - L(\tau)\}$ *converges weakly to a zero-mean Gaussian process.*

The detailed proof of Theorem 2 and the form of the influence function $\pi_i(\tau, \tilde{\boldsymbol{\beta}})$ are deferred to Supplementary Materials A.2. We further establish in Supplementary Materials A.3 that the distribution of $\sqrt{n}\{\hat{L}_{CV}(\tau, \hat{\boldsymbol{\beta}}) - L(\tau)\}$ is asymptotically equivalent to that of

$\sqrt{n}\{\hat{L}_n(\tau, \hat{\boldsymbol{\beta}}) - L(\tau)\}$. As a result, $\sqrt{n}\{\hat{R}^1_n(\tau, \hat{\boldsymbol{\beta}}) - R^1(\tau)\}$ and $\sqrt{n}\{\hat{R}^1_{CV}(\tau, \hat{\boldsymbol{\beta}}) - R^1(\tau)\}$ are also asymptotically equivalent and converge weakly to a zero-mean Gaussian process. Their influence functions are provided in Supplementary Materials A.2.

For covariance and interval estimation, we also develop a perturbation resampling scheme similar to those in Jin et al. (2001). We consider a perturbed objective function, defined as

$$\hat{L}_n^*(\tau, \boldsymbol{b}) = n^{-1}\sum_{i=1}^n \omega_i \frac{\Delta_i}{G^*(Y_i^u)}\rho_\tau(Y_i^u - \boldsymbol{Z}_i\boldsymbol{b})$$

where $\{\omega_i\}_{i=1}^n$ are i.i.d. unit exponential random variates. Here $G^*(t)$ is a perturbed version of the Kaplan-Meier estimator using the same set of weights. That is, $G^*(t) = \mathscr{P}_{s\in[0,t]}\{1 - dN^*(t)/Y^*(t)\}$, where $N^*(t) = n^{-1}\sum_{i=1}^n \omega_i I(Y_i \le t, \delta_i = 1)$ and $Y^*(t) = n^{-1}\sum_{i=1}^n \omega_i I(Y_i \ge t)$, and $\mathscr{P}$ is the product-integral operator. Define $\hat{\boldsymbol{\beta}}^*(\tau)$ as the minimizer of $\hat{L}_n^*(\tau, \boldsymbol{b})$ with respect to $\boldsymbol{b}$. In Supplementary Materials A.4, we show that the unconditional distribution of $\sqrt{n}\{\hat{L}_n(\tau, \hat{\boldsymbol{\beta}}) - L(\tau)\}$ is asymptotically the same as the conditional distribution of $\sqrt{n}\{\hat{L}_n^*(\tau, \hat{\boldsymbol{\beta}}^*) - \hat{L}_n(\tau, \hat{\boldsymbol{\beta}})\}$. In practice, one may generate B perturbed samples of $\hat{L}_n^*(\tau, \hat{\boldsymbol{\beta}}^*)$ for a large number B, denoted by $\{\hat{L}_n^{*b}(\tau, \hat{\boldsymbol{\beta}}^*)\}_{b=1}^B$. The standard error estimates for the proposed estimators of $L(\tau)$ and the confidence intervals for $L(\tau)$ can be obtained from the empirical standard deviations and the empirical percentiles of $\sqrt{n}\{\hat{L}_n^{*b}(\tau, \hat{\boldsymbol{\beta}}^*) - \hat{L}_n(\tau, \hat{\boldsymbol{\beta}})\}$.

4. THE PROPOSED MODEL COMPARISON PROCEDURES

In practice, there are often more than one candidate models under consideration. Based on the arguments in Section 2, we can formally compare the predictive performance between two working models by comparing their respective $L(\tau)$'s. Let $\boldsymbol{Z}_A$ and $\boldsymbol{Z}_B$ be two subvectors of $\bar{\boldsymbol{Z}}$ and both of them include 1 as the first element. The working model (2) with $\boldsymbol{Z} = \boldsymbol{Z}_A$ and the working model (2) with $\boldsymbol{Z} = \boldsymbol{Z}_B$ are referred to as model A and model B, respectively. Their corresponding prediction loss $L(\tau)$'s are denoted by $L_A(\tau)$ and $L_B(\tau)$. Likewise, in the sequel, subscripts, such as A and B, will be used to differentiate quantities for different working models. From the arguments in Section 2, $L_A(\tau) - L_B(\tau)$ can be used to capture the relative prediction loss of two *working* models at the τ_{th} quantile level. This motivates us to compare the predictive capacity of two candidate models based on their difference in $L(\tau)$.

4.1 Model comparisons at a fixed τ

When the interest lies in a single quantile at level τ, we formulate the predictive capacity comparison as a hypothesis testing problem regarding $L_A(\tau) - L_B(\tau)$, which, by Theorem 1, can be consistently estimated by $\mathscr{T}_{AB}(\tau) \equiv \hat{L}_{nA}(\tau, \hat{\boldsymbol{\beta}}_A) - \hat{L}_{nB}(\tau, \hat{\boldsymbol{\beta}}_B)$. However, a notable complication is that the limiting distribution of $\mathscr{T}_{AB}(\tau)$ under the null, $H_0: L_A(\tau) = L_B(\tau)$, is different under the nested model scenario and the non-nested model scenario. Thus, the hypothesis testing method needs to be designed differently for these two scenarios.

When the two models are non-nested, in the sense that $\boldsymbol{Z}_A$ contains covariates not contained in $\boldsymbol{Z}_B$ and vice versa, we consider the following two-sided hypothesis testing problem:

$$H_0: L_A(\tau) = L_B(\tau) \quad \text{vs.} \quad H_1^{nn}: L_A(\tau) \ne L_B(\tau), \quad (9)$$

where H_0 corresponds to the scenario when models A and B have equivalent predictive capacity for the τ_{th} quantile. Naturally, we adopt $\mathscr{T}_{AB}(\tau)$ as the test statistics. Based on Theorem 1, we have

$$\mathscr{T}_{AB}(\tau) - \{L_A(\tau) - L_B(\tau)\} = n^{-1}\sum_{i=1}^{n}\{\pi_{iA}(\tau, \tilde{\beta}_A) - \pi_{iB}(\tau, \tilde{\beta}_B)\} + o_p(n^{-1/2}). \tag{10}$$

From the definition of $\pi_i(\tau, \tilde{\beta})$ in Supplementary Materials A.2., we see that $\pi_{iA}(\tau, \tilde{\beta}_A) - \pi_{iB}(\tau, \tilde{\beta}_B)$ is not degenerate provided $Z_A\tilde{\beta}_A(\tau) \not\equiv Z_B\tilde{\beta}_B(\tau)$, which is a reasonable assumption for the non-nested case. When this holds, we can show that $\mathscr{T}_{AB}(\tau)$ is of order $O_p(n^{-1/2})$ under H_0, and that $\sqrt{n}\mathscr{T}_{AB}(\tau)$ converges to a zero-mean Normal distribution. We employ a resampling procedure to approximate the limit distribution of $\mathscr{T}_{AB}(\tau)$ under H_0. Let

$$\mathscr{T}_{AB}^{nn*}(\tau) - \{\hat{L}_{nA}^{*}(\tau, \hat{\beta}_A^{*}) - \hat{L}_{nA}(\tau, \hat{\beta}_A)\} - \{\hat{L}_{nB}^{*}(\tau, \hat{\beta}_B^{*}) - \hat{L}_{nB}(\tau, \hat{\beta}_B)\},$$

where ^{nn} stands for non-nested. By the arguments in Supplementary Materials A.4, we see that the conditional distribution of $\sqrt{n}\mathscr{T}_{AB}^{nn*}$ given the observed data is asymptotically equivalent to the unconditional distribution of $\sqrt{n}\mathscr{T}_{AB}$ under H_0. Therefore, one may obtain distribution cut-off values for $\mathscr{T}_{AB}(\tau)$ based on its perturbed samples, $\{|\mathscr{T}_{AB}^{nn*b}(\tau)|\}_{b=1}^{B}$, and reject H_0 when $|\mathscr{T}_{AB}(\tau)|$ is greater than the cut-off value.

We next consider the more complicated scenario that involves two nested models. Suppose Z_A is a subvector of Z_B and thus $L_B(\tau)$ ≤ $L_A(\tau)$. We assume that the elements in Z_B are linearly independent with each other, in the sense that there does not exist a non-zero vector $\mathfrak{a}$ satisfying $Z_B\mathfrak{a} \equiv 0$. We formulate a one-sided hypothesis test,

$$H_0 : L_A(\tau) = L_B(\tau) \quad \text{vs.} \quad H_1^n : L_B(\tau) < L_A(\tau). \tag{11}$$

Rejecting H_0 would imply a significant improvement by adopting model B (versus model A) in predicting the τth-quantile, which may lead to a decision of adopting model B instead of model A in constructing the quantile predictor. In the nested case, H_0 corresponds to the situation when $Z_A\tilde{\beta}_A \equiv Z_B\tilde{\beta}_B$. Note that the asymptotic representation in (10) is not useful, because $\pi_{iA}(\tau, \tilde{\beta}_A) - \pi_{iB}(\tau, \tilde{\beta}_B) \equiv 0$. As a result, $\mathscr{T}_{AB}(\tau)$ is no longer of order $O_p(n^{-1/2})$ under H_0 but has a faster convergence rate to 0. The faster convergence to 0 was also noted in Demler et al. (2012) for comparing area under curve between nested models.

To address this difficulty, we develop a different perturbation scheme by extending the method in Chen et al. (2008). Specifically, we propose the following perturbed test statistics,

$$\mathscr{T}_{AB}^{n*}(\tau) = \{\hat{L}_{nA}^{*}(\tau, \hat{\beta}_A^{*}) - \hat{L}_{nA}^{*}(\tau, \hat{\beta}_A)\} - \{\hat{L}_{nB}^{*}(\tau, \hat{\beta}_B^{*}) - \hat{L}_{nA}^{*}(\tau, \hat{\beta}_B)\},$$

where ^{n} stands for nested. Note that $\mathcal{T}_{AB}^{n*}(\tau)$ differs from $\mathcal{T}_{AB}^{nn*}(\tau)$ by including $\hat{L}_{nA}^{*}(\tau,\hat{\beta}_{A})$ and $\hat{L}_{nB}^{*}(\tau,\hat{\beta}_{B})$ as the second and fourth term, in place of $\hat{L}_{nA}(\tau,\hat{\beta}_{nA})$ and $\hat{L}_{nB}(\tau,\hat{\beta}_{nB})$. This minor modification allows us to achieve a fine enough approximations to $\mathcal{T}_{AB}(\tau)$, which is of order $O_p(n^{-1})$ under H_0. Specifically, Theorem 3 provides the crucial result for justifying the use of $\mathcal{T}_{AB}^{n*}(\tau)$. Our proof for this theorem in Supplementary Materials B.1. follows the lines of Rao and Zhao (1992)(see Lemma 2.2) and Chen et al. (2008).

Theorem 3—*Define* $J(\tau)=\dfrac{\partial}{\partial b}E\{ZI(T \leq Z'b)\}|_{b=\tilde{\beta}(\tau)}.$ *Under conditions* **C1–C4**,

$$n\{\hat{L}_n(\tau,\hat{\beta}) - \hat{L}_n(\tau,\tilde{\beta})\}= -\frac{n}{2}\{\hat{\beta}(\tau) - \tilde{\beta}(\tau)\}'J(\tau)\{\hat{\beta}(\tau) - \tilde{\beta}(\tau)\}+o_p(1); \quad (12)$$

Similarly, for the perturbed counterpart of the loss function, we have

$$n\{\hat{L}_n^{*}(\tau,\hat{\beta}^{*}) - \hat{L}_n^{*}(\tau,\hat{\beta})\}= -\frac{n}{2}\{\hat{\beta}^{*}(\tau) - \hat{\beta}(\tau)\}'J(\tau)\{\hat{\beta}^{*}(\tau) - \hat{\beta}(\tau)\}+o_p(1). \quad (13)$$

It is worth noting that the $J(\tau)$ matrix equals $E[Z^{\otimes 2}f\{Z\tilde{\beta}(\tau)|Z\}]$. Due to the unknown conditional density f, it is rather hard to directly estimate matrix $J(\tau)$ in practice. Based on Theorem 3, we derived the asymptotic distribution of $n\mathcal{T}_{AB}(\tau)$ and the limit conditional distribution of $n\mathcal{T}_{AB}^{n*}(\tau)$ given the observed data; see Supplementary Materials B.2. Our results justify the use of $\mathcal{T}_{AB}^{n*}$ for approximating the distribution of $\mathcal{T}_{AB}^{n*}(\tau)$.

Remark—Theorem 3 also allows us to correct for the small bias in $\hat{L}_n(\tau,\hat{\beta})$. From (12), we see that the bias in $\hat{L}_n(\tau,\tilde{\beta})$ due to using the entire observed dataset twice is asymptotically

$-\dfrac{1}{2}\{\hat{\beta}(\tau) - \tilde{\beta}(\tau)\}'J(\tau)\{\hat{\beta}(\tau) - \tilde{\beta}(\tau)\}$, which is non-positive and of order $O_p(n^{-1})$. According to (13), we propose to use

$$\hat{L}_{\text{adj}}(\tau,\hat{\beta})=\hat{L}_n(\tau,\hat{\beta}) - B^{-1}\sum_{b=1}^{B}\{\hat{L}_n^{*b}(\tau,\hat{\beta}^{*}) - \hat{L}_n^{*b}(\tau,\hat{\beta})\},$$

where B is the total number of perturbations. The adjustment to $\hat{L}_n(\tau,\hat{\beta})$ is of order $O_p(n^{-1})$; thus the influence functions for $\hat{L}_n(\tau,\hat{\beta})$, $\pi_A(\tau,\tilde{\beta})$, still apply to $\hat{L}_{adj}(\tau,\hat{\beta})$. This adjustment can also be implemented in the test statistics $\mathcal{T}_{AB}(\tau)$ when testing non-nested models, but not in the comparison of nested models.

4.2 Model comparisons across a range of τ

In practice, it is often desirable to compare two models by accounting for a range of quantiles, so as to identify an optimal model for predicting a range of quantiles simultaneously. A model comparison procedure that addresses a range of quantiles simultaneously can be achieved through the R^1 measure introduced in Section 2.3.

Specifically, we propose to compare two candidate models, namely model A and B, by testing the hypotheses, $H_0^{nn}: R_A^1 = R_B^1$ vs. $H_1: R_A^1 \neq R_B^1$ when model A and model B are non-nested, and the hypothesis $H_0^n: R_A^1 = R_B^1$ vs. $H_1: R_A^1 < R_B^1$ in the nested case where $\mathbf{Z}_A$ is a subvector of $\mathbf{Z}_B$.

Next, we notice that $R_A^1 - R_B^1 = \dfrac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} \dfrac{L_B(\tau) - L_A(\tau)}{E\rho_\tau(T^{0u} - \zeta_\tau)} d\tau$. This motivates us to define

$$\mathscr{R}_{AB} = \frac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} \frac{\hat{L}_{nB}(\tau, \hat{\boldsymbol{\beta}}_A) - \hat{L}_{nA}(\tau, \hat{\boldsymbol{\beta}}_B)}{\hat{L}_n(\tau, \hat{\zeta}_\tau)} d\tau. \tag{14}$$

Essentially, $\mathscr{R}_{AB}$ is a weighted average of $\mathscr{T}_{AB}$ across $\boldsymbol{\tau}$. The distribution of $\mathscr{R}_{AB}$ under H_0 can be approximated through the empirical distribution of

$\mathscr{R}_{AB}^{n*} = -\dfrac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} \{ \mathscr{T}_{AB}^{n*} / \hat{L}_n^*(\tau, \hat{\zeta}_\tau^*) \} d\tau$ in the nested case, and through

$\mathscr{R}_{AB}^{nn*} = -\dfrac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} \{ \mathscr{T}_{AB}^{nn*} / \hat{L}_n^*(\tau, \hat{\zeta}_\tau^*) \} d\tau$ in the non-nested case.

5. SIMULATIONS

5.1 Estimation of the prediction performance measures

Simulations were conducted to assess the finite-sample performances of the proposed methods. The first part of our simulations illustrates the proposed measures, $L(\tau)$ and $R^1(\tau)$, under various working prediction models, and furthermore examines whether they can be consistently estimated by the proposed estimators. We generated the survival times as

$$T^* = \exp\{2 \times Z_{10} + I(Z_2=1)\varepsilon_1 + I(Z_2=0)\varepsilon_2 + Z_3 + \varepsilon_3\},$$

where Z_{10} follows a truncated normal distribution with mean 0, standard deviation 0.5, and truncation points ± 1.5. The variable Z_2 follows a Bernoulli distribution with expectation 0.5, and $Z_3 \sim Uniform(-0.5, 0.5)$. The ε_1, ε_2, and ε_3 are mutually independent and follow $Normal(0, 1^2)$, $Normal(0, 0.2^2)$ and $Normal(0, 0.25^2)$ respectively. Therefore, we have

$$Q_T(\tau | \mathbf{Z}) = q_{\varepsilon 2 + \varepsilon 3}(\tau) + 2 \times Z_{10} + \{ q_{\varepsilon 1 + \varepsilon 3}(\tau) - q_{\varepsilon 2 + \varepsilon 3}(\tau) \} \times Z_2 + Z_3. \tag{15}$$

The coefficient of Z_2 is a monotone function of τ that crosses 0 at $\tau = 0.5$. The censoring time $C^* = \exp(C)$, where $C = \zeta_C \times runif(-1.2, 2.5) + (1 - \zeta_C) \times 2.5$ and $\zeta_C \sim Bernoulli(0.8)$. The censoring rate is 28%, and we set $u = 2.49$.

Our goal is to assess the predictive capacity of several working models, namely (A) $Z_{10} + Z_2 + Z_3$, (B) $Z_1 + Z_2 + Z_3$, (C) $Z_{10} + Z_3$, (D) $Z_1 + Z_3$, and (E) $Z_{10} + Z_2 + Z_3 + Z_4 + Z_5 + Z_6$. Here, each working model is represented by an addition sentence, the terms of which indicate the covariates included in $\mathbf{Z}$. Each model can facilitate a quantile predictor $\hat{\xi}_\tau(\cdot)$.

The covariate $Z_1 = Z_{10} + \varepsilon_4$, where $\varepsilon_4 \sim Uniform(-0.25, 0.25)$. Thus Z_1 represents a mis-measured version of Z_{10}. Also, $Z_4 \sim Uniform(-1, 1)$, $Z_5 = \text{sign}(Z_{10}) \cdot \sqrt{|Z_{10}|}$, and $Z_6 \sim 2 \times Beta(2, 2)$. Therefore, Model (A) represents the data generation model in (15). Model (B) corresponds to the situation where one covariate in the true model is subject to mis-measurement. In Model (C), one leaves out a covariate that explains the heteroscedasticity of the error terms. Model (D) is another mis-specified model that includes Z_1 instead of Z_{10} and excludes Z_2. Finally, model (E) includes several redundant covariates. It exemplifies a working prediction model that is over-fitted and has the problem of collinearity.

To visualize the corresponding prediction loss and performance measures $L(\tau)$ and $R^1(\tau)$ under (A)–(E), we plot their theoretical values, which were obtained via large-sample approximations based on 10, 000 simulated uncensored data points. We observe in Figure 1 that the two measures effectively describe the predictive capacity of the working models. For example, we observe from the left panel of Figure 1 that model (A) and (E) always render the smallest $L(\tau)$. Model (B) leads to larger prediction loss, by including a mis-measured covariate Z_1 instead of the true Z_{10}. Model (C) yields the same prediction loss as (A) and (E) when $\tau = 0.5$, for which the conditional quantile does not depend on Z_2. However, the quantile prediction is generally worse at other quantile levels. Model (D) is worse than (B) and (C) by deviating further away from the true model. In addition, we observe from the right panel of Figure 1 that $R^1(\tau)$ transforms the prediction loss into a predictive performance measure of unified scales, thereby facilitating comparisons across a range of τ.

We examined whether the proposed methods can provide adequate estimation and inference for the prediction loss/performance measures. For each working model, we conducted 2, 000 simulations under sample sizes $n = 200, 400$ and 600. We present the simulation results under model (A) and (B) in Table 1 for $\tau = 0.1, 0.3, 0.5$ and 0.6. The results under other working models are similar and thus are not reported. The reported summary statistics include the empirical bias (EB), the empirical standard error (ESE), the average of influence function based standard error (ASE), and empirical coverage probabilities of 95% Wald-type confidence intervals (C95). To build a confidence interval for $L(\tau)$, we used a log transformation and the Delta method. The $\log\{-\log(x)\}$ transformation was used to build a confidence interval for $R^1(\tau)$. The performance of the bias-adjusted estimators in Section 4.1 is also summarized. Table 1 suggests that $L(\tau)$ and $R^1(\tau)$ are estimated accurately by the proposed plug-in estimators $\hat{L}_n(\tau, \hat{\boldsymbol{\beta}})$ and $\hat{R}_n^1(\tau, \hat{\beta})$. The standard errors are generally small. The coverage rates of the confidence intervals for $L(\tau)$ are slightly low when $n = 200$ but get closer to the nominal level for larger n's. The bias-adjusted methods show some improvement when $n = 200$, but the improvement become negligible when $n \geq 400$. The performances of the CV-estimators are similar and deferred to Supplementary Materials C, Table C.1.

We also conducted sensitivity studies by setting $C = \log\{2.05\varpi^*\exp(0.25 \times Z_2 + 0.25 \times Z_3)\}$, where ϖ follow exponential distribution with rate 0.5. The censoring time $\exp(C)$ thus follows a Cox proportional hazards model, and the censoring rate remains 28%. We implemented the proposed methods with $\hat{G}(\cdot)$ adopted, pretending C is still independent of $\mathbf{Z}$. Under such a mis-specification of the censoring distribution, we observe from Table C.2

in supplemental material C that the bias of $\hat{R}_n^1(\tau, \hat{\beta})$ are still very small. The empirical coverage rates may be slightly lower than the nominal level but are still reasonable.

5.2 Inferences on predictive capacity

In the second part of our simulations, we examine the utility of the proposed hypothesis testing method for comparing the quantile predictive performance between candidate prediction models, with the true predictive performance measures displayed in Figure 1. We considered five scenarios: (i) A vs. B, (ii) A vs. C, (iii) E vs. A, (iv) B vs. C, and (v) B vs. D. The one-sided hypothesis was adopted for evaluating (ii), (iii) and (v), while two-sided hypothesis was adopted for (i) and (iv). We considered two significance levels, $\alpha = 0.05$ and $\alpha = 0.1$, and set re-sampling sample size $B = 1,999$. The results were summarized in Table 2. We observe that the empirical sizes are generally close to the nominal levels, no matter whether the working models under comparison are correctly specified or not. Furthermore, the empirical powers appear satisfactory. The influence functions we derive also facilitate Wald-type confidence intervals for the difference in $R^1(\tau)$ between two non-nested models. Some simulation results were provided in Table C.3 of Supplemental Material C.

We next conducted model comparisons for $\tau \in [0.1, 0.6]$ based on $\mathcal{R}_{AB}$, the overall test statistic. Taking all quantiles into consideration, each pair of compared models do not have the same predictive performance except under scenario (iii), E vs. A. The results in Table C. 4 of Supplementary Material C suggest that the hypothesis testing procedures maintained the sizes well when H_0 is true. The empirical power increases with the sample size as well as the expected model discrepancy. We also conducted sensitivity analysis to evaluate the hypothesis testing procedure under the aforementioned covariate-dependent censoring scenario, where $\hat{G}(\cdot)$ is obtained using the Kaplan-Meier estimator. From Table C.5 of Supplementary Material C, we observe that the hypothesis testing procedure still performs adequately under moderate violations of the covariate-independent censoring assumption.

6. DATA ANALYSIS

We applied the proposed methods to a renal disease study, which examined dialysis mortality in a cohort of incident dialysis patients aged 20 years and older (Kutner et al., 2002). The primary endpoint was time to death since study enrollment, and the mean follow-up time was approximately 3 years. The mean age at baseline was 56 years. Covariates include severity of restless syndrome (LEGS), age in years (AGE), BMI, dialysis mordality (HDPD), hematocrit (HCT), mental health score (MH), ferritin level (FER), albumin level (ALB), primary diagnosis of diabetes at study start (PRIDIAB), presence of cardiovascular comorbidity (NEWCAR), higher education level (HIEDU), male sex (MALE), fish consumption (FISHH), and black race (BLACK) (Peng and Huang, 2008; Peng et al., 2014). The survival endpoint may be censored by renal transplantation or end of study follow-up, leading to an overall censoring rate of 35%. There were 191 observations with complete data in this dataset.

Our goal here is to evaluate and compare the predictive capacity of different *working* quantile regression models. Our preliminary results based on the proposed methods

suggested two potential candidates, namely Model A: {AGE, BLACK, FISHH, HCT, LEGS}, and Model B: {AGE, BLACK, FISHH, HCT}. To have a more comprehensive illustration, we also considered several alternative models here, some of which have been studied in literature (Peng et al., 2014). See Table 3 for details. Specifically, Models B–F are sub-models of model A that includes one less covariate. In model G, AGE_B is a binary variable that indicates whether the patient was greater than or equal to 60 years of age. This is to illustrate whether the common practice of stratefying age makes any difference in this data analysis. Model H denotes the model selected by the adaptive-LASSO method under the accelerated failure time (AFT) model (Wang and Leng, 2007). Finally, Model I includes two additional covariates besides those in Model A. The overall prediction performance measure $\hat{R}^1 = \frac{1}{\tau_U - \tau_L} \int_{\tau_L}^{\tau_U} \hat{R}_n^1(\tau, \hat{\beta}) d\tau$ for $[\tau_L, \tau_U] = [0.05, 0.65)$ was also provided in Table 3. We plotted the pointwise $\hat{R}_n^1(\tau, \hat{\beta})$ for $\tau \in [0.05, 0.65)$ in Figure 2. Table 4 displays the results of the formal hypothesis tests, either overall or at specific τ 's.

The left panel of Figure 2 and the top section of Table 4 compare model A to its several sub-models, models B–F. These results reveal the relative importance of each covariate. For example, we observe that dropping AGE from Model A resulted in the biggest decline in $R^1(\tau)$, suggesting that AGE is crucial for quantile prediction across τ. This is confirmed by the hypothesis testing results, where the p-values for comparing Model A vs. D were almost always < .001. By comparison, the decline in $R^1(\tau)$ by dropping LEGS from model A appears larger for smaller τ 's but diminishes when τ increases. Coupled with the p-values for comparing Model A vs. B, this result may suggest that LEGS is non-negligible for lower quantiles but may be less important for upper quantiles. This finding is consistent with previous analysis of this dataset (Peng and Huang, 2008). Thus, one may prefer Model B when research interest lies only in the higher quantiles but may favor Model A when the lower quantiles, corresponding to the higher risks patients, are also of interest.

In the right panel of Figure 2 and the bottom session of Table 4, we illustrate more scenarios of model comparisons. When comparing Model A with Model G, we observe that the continuous AGE covariate served significantly better as a predictor, as compared to its dichotomized counterpart. Note that the proposed methods allow us to evaluate the influence of covariate transformations on quantile prediction. Next, we observe that the $\hat{R}_n^1(\tau, \hat{\beta})$ of Model A generally lie above that of Model H, the model selected by the adaptive Lasso method under AFT model. Although the difference between Models A and H are not statistically significant, Model A may be preferred by rendering higher $\hat{R}_n^1(\tau, \hat{\beta})$ with the same number of covariates. Finally, we observe that adding additional predictors, HDPD and HIEDU, into Model A does not lead to significant improvement in prediction performances. This bigger model, Model I, did outperformed Model B at lower quantiles, possibly by including LEGS.

Through these visual and formal comparisons, we conclude that Model A may be adopted if one wishes to achieve good quantile prediction for both the lower quantiles and the higher quantiles. Alternatively, Model B may be preferred in terms of overall prediction performance. The regression coefficient estimates under Model A are presented in

Supplemental Material C. Using this data example, we demonstrate that the proposed methods allow us to evaluate the quantile prediction performance of various working models. It also facilitates comprehensive and rigorous comparisons in predictive capacity between various candidate models.

7. DISCUSSION

In this work, we have proposed a rigorous framework for evaluating the predictive capacity of *working* quantile regression models in predicting the quantiles of a survival outcome. The proposed framework allows for not only quantification of either local or global predictive performances, but also formal comparisons between two models that are either nested or non-nested. By its robustness to model mis-specification and straightforward interpretation, we believe that the proposed framework features high practical utility in biomedical studies.

The proposed framework bears a fundamental difference from the existing goodness-of-fit methods by targeting the performance in predicting the underlying survival quantiles of *new* patients. In contrast, goodness-of-fit methods typically evaluate how the observed *existing* data match the data expected under an assumed model. The key distinction is well reflected by the construction of the proposed prediction loss function, $E[\rho_\tau\{T^0 - \xi_\tau(Z^0)\}]$, which takes expectation over the data obtained from a *new* subject, (T^0, Z^0), rather than the data where the quantile prediction rule ξ_τ is derived from. Our proposals in Section 4 demonstrate a useful application of the proposed framework. That is, they provide formal inference procedures for conducting model comparison/selection based on the criterion formulated according to the proposed predictive performance measure.

The proposed framework is general and allows one to adopt other forms of quantile predictors. Employing a different $\hat{\xi}_\tau(\cdot)$ would only require moderate modifications to the inference procedures. One may need to adapt the components involving the variability of $\hat{\xi}_\tau(\cdot)$ in the derivation of the influence function $\pi_j(\tau)$. The asymptotic null distribution for the test of nested models would take a different form but can be derived by similar lines.

The proposed $L(\tau)$ is constructed based on log-transformed survival time T. Given the equivariance property of quantiles, which implies $Q_T(\tau|\bar{Z}) = \log\{Q_{\exp(T)}(\tau|\bar{Z})\}$, $L(\tau)$ can be interpreted as a measure capturing the discrepancy in log survival quantiles. Note that censored quantile regression typically models a log-transformed survival time, and the resulting quantile predictions often present a skewed distribution across subjects. Assessing quantile prediction performance in the log-time scale can protect against such skewness.

In Section 2.1, one may want to rescale the summands of $L_n(\tau, \hat{\beta})$ by the inverse of the conditional density $f(\xi_\tau^0|\bar{Z})$. However, direct estimation of $f(\xi_\tau^0|\bar{Z})$ typically involves smoothing and can be unstable with small to moderate sample size. In practice, we recommend using $\hat{R}_n^1(\tau, \hat{\beta})$, a scaled alternative to $L_n(\tau, \hat{\beta})$ which is also easy to interpret.

The proposed procedures assume that p is finite. When the number of covariates in a real dataset is large, one may first apply existing variable selection procedures for quantile regression models (Wu and Liu 2009; Li and Zhu 2008; Zou and Yuan 2008; Peng et al.

2014; among others) to reduce the model dimensionality. Our proposals in this work can provide an objective analytical tool to validate and compare the models resulted from various different selection procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Angrist J, Chernozhukov V, Fernández-Val I. Quantile regression under misspecification, with an application to the US wage structure. Econometrica. 2006; 74:539–563.

Chen K, Ying Z, Zhang H, Zhao L. Analysis of least absolute deviation. Biometrika. 2008; 95:107–122.

Chen PY, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics. 2001; 57:1030–1038. [PubMed: 11764241]

Demler OV, Pencina MJ, D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. Statistics in medicine. 2012; 31:2577–2587. [PubMed: 22415937]

Goldberg Y, Kosorok MR. Q-learning with censored data. The Annals of Statistics. 2012; 40:529–560. [PubMed: 22754029]

He X, Zhu LX. A lack-of-fit test for quantile regression. Journal of the American Statistical Association. 2003; 98:1013–1022.

Huang Y. Quantile calculus and censored regression. The Annals of Statistics. 2010; 38:1607–1637. [PubMed: 20592942]

Jin Z, Ying Z, Wei L. A simple resampling method by perturbing the minimand. Biometrika. 2001; 88:381–390.

Koenker, R. Quantile regression. Vol. 38. Cambridge Univ Pr; 2005.

Koenker R, Bassett G. Regression quantiles. Econometrica. 1978; 46:33–50.

Koenker R, Machado J. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association. 1999; 94:1296–1310.

Kutner NG, Clow PW, Zhang R, Aviles X. Association of fish intake and survival in a cohort of incident dialysis patients. American Journal of Kidney Diseases. 2002; 39:1018–1024. [PubMed: 11979345]

Lawless JF, Yuan Y. Estimation of prediction error for survival models. Statistics in Medicine. 2010; 29:262–274. [PubMed: 19882678]

Li Y, Zhu J. L1-norm quantile regression. Journal of Computational and Graphical Statistics. 2008; 17:163–185.

Noh H, El Ghouch A, Van Keilegom I. Assessing model adequacy in possibly misspecified quantile regression. Computational Statistics & Data Analysis. 2013; 57:558–569.

Peng L, Huang Y. Survival analysis with quantile regression models. Journal of the American Statistical Association. 2008; 103:637–649.

Peng L, Xu J, Kutner N. Shrinkage estimation of varying covariate effects based on quantile regression. Statistics and Computing. 2014; 24:853–869. [PubMed: 25332515]

Portnoy S. Censored regression quantiles. Journal of the American Statistical Association. 2003; 98:1001–1012.

Rao CR, Zhao L. Approximation to the distribution of m-estimates in linear models by randomly weighted bootstrap. Sankhy : The Indian Journal of Statistics, Series A. 1992; 54:323–331.

Tian L, Cai T, Goetghebeur E, Wei L. Model evaluation based on the sampling distribution of estimated absolute prediction error. Biometrika. 2007; 94:297–311.

Tian L, Zhao L, Wei L. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014; 15:222–233. [PubMed: 24292992]

Uno H, Cai T, Tian L, andWei L. Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association. 2007; 102:527–537.

Wang H, Leng C. Unified LASSO estimation by least squares approximation. Journal of the American Statistical Association. 2007; 102:1039–1048.

Wang H, Wang L. Locally weighted censored quantile regression. Journal of the American Statistical Association. 2009; 104:1117–1128.

Wang L. Nonparametric test for checking lack of fit of the quantile regression model under random censoring. Canadian Journal of Statistics. 2008; 36:321–336.

Wu Y, Liu Y. Variable selection in quantile regression. Statistica Sinica. 2009; 19:801–817.

Ying Z, Jung S, Wei L. Survival analysis with median regression models. Journal of the American Statistical Association. 1995; 90:178–184.

Zhou L. A simple censored median regression estimator. Statistica Sinica. 2006; 16:1043–1058.

Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. The Annals of Statistics. 2008; 36:1108–1126.

Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. Journal of the American Statistical Association. 1998; 93:702–709.
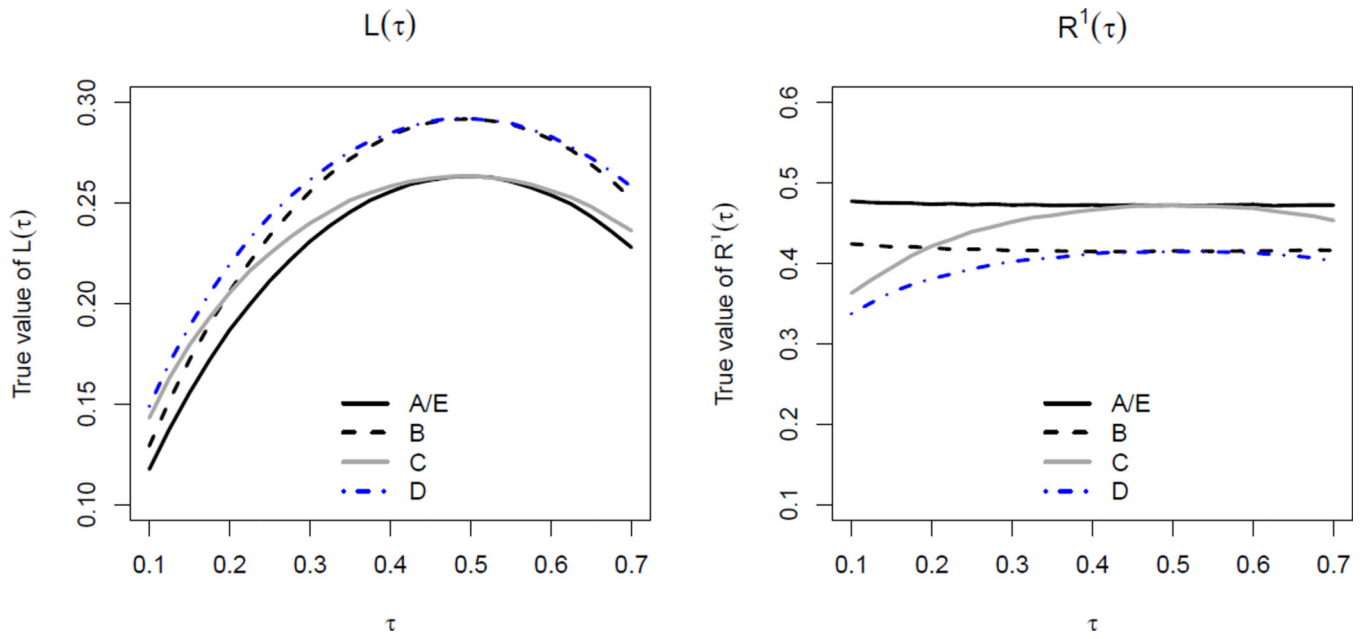
Figure 1.
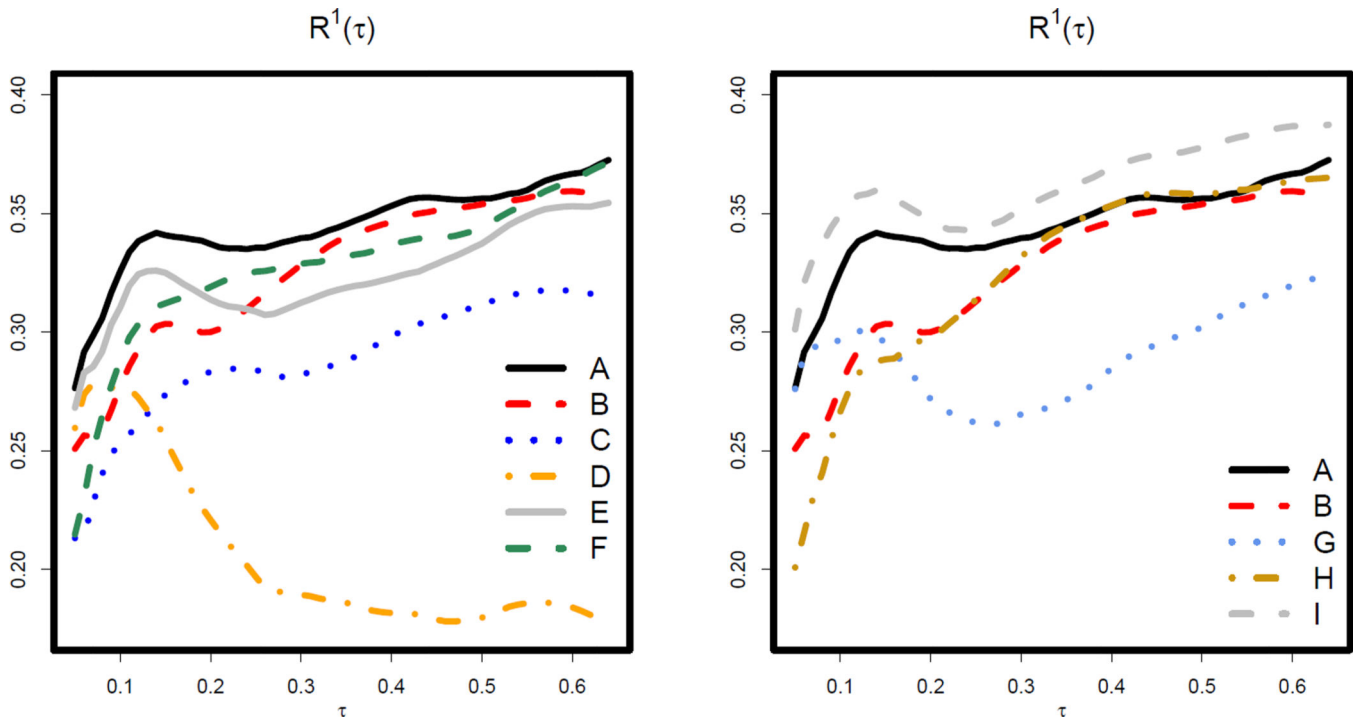Values of $L(\tau)$ and $R^1(\tau)$ under different working models.

Figure 2.

Analysis of the dialysis data: $\hat{R}_n^1(\tau, \hat{\beta})$ under Models A–I.

Table 1

Simulations: summary statistics for $\hat{L}_n(\tau, \hat{\beta})$ and $\hat{R}_n^1(\tau, \hat{\beta})$ at $\tau = 0.1, 0.3, 0.5, 0.6$ under model A and model B. The expected number of events $E(\sum_{i=1}^n \delta_i)$ equal 144, 288 and 432 when $n = 200$, 400 and 600 respectively. The results of the bias-adjusted estimators are also included.

n	τ	Model A					Model B				
		TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)
						$\hat{L}_n(\tau, \hat{\beta})$					
200	0.1	0.117	−5	11	11	90.4	0.129	−5	11	11	90.6
	0.3	0.231	−6	21	20	92.3	0.255	−7	21	21	92.4
	0.5	0.263	−7	25	24	91.8	0.291	−8	25	24	91.3
	0.6	0.253	−7	25	23	91.0	0.281	−7	25	24	90.6
400	0.1	0.117	−2	8	8	93.2	0.129	−2	8	8	93.2
	0.3	0.231	−2	15	15	93.5	0.255	−3	15	15	93.9
	0.5	0.263	−3	18	17	92.9	0.291	−4	18	17	93.5
	0.6	0.253	−3	18	17	92.6	0.281	−4	18	17	93.0
600	0.1	0.117	−1	7	6	93.7	0.129	−1	7	7	93.9
	0.3	0.231	−1	12	12	94.6	0.255	−2	12	12	94.0
	0.5	0.263	−2	14	14	94.2	0.291	−2	14	14	93.6
	0.6	0.253	−2	14	14	93.6	0.281	−2	14	14	93.2
						$\hat{R}_n^1(\tau, \hat{\beta})$					
200	0.1	0.478	11	52	51	93.3	0.425	11	53	52	93.4
	0.3	0.473	7	48	47	93.9	0.417	8	49	48	94.4
	0.5	0.472	7	51	49	93.2	0.415	8	51	50	93.8
	0.6	0.473	7	53	51	93.1	0.416	8	55	52	93.0
400	0.1	0.478	6	37	36	93.7	0.425	6	37	37	93.8
	0.3	0.473	4	34	33	93.7	0.417	5	34	34	93.8
	0.5	0.472	5	36	35	93.8	0.415	6	36	35	93.8
	0.6	0.473	5	37	36	93.5	0.416	6	38	37	93.6

n	τ	Model A					Model B				
		TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)
600	0.1	0.478	5	30	29	94.1	0.425	4	30	30	94.2
	0.3	0.473	3	28	27	94.3	0.417	4	28	28	94.6
	0.5	0.472	3	29	28	94.0	0.415	4	29	29	94.4
	0.6	0.473	4	31	30	93.8	0.416	4	31	30	94.1

$\hat{L}_{adj}(\tau, \hat{\beta})$

n	τ	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)
200	0.1	0.117	−1	12	11	93.0	0.129	−1	12	11	92.4
	0.3	0.231	−2	21	21	93.9	0.255	−2	22	21	93.5
	0.5	0.263	−2	25	24	93.4	0.291	−2	25	24	93.2
	0.6	0.253	−2	25	23	92.4	0.281	−2	25	24	92.9
400	0.1	0.117	0	8	8	93.3	0.129	0	8	8	93.6
	0.3	0.231	−1	15	15	93.5	0.255	−1	15	15	93.9
	0.5	0.263	−1	18	17	93.1	0.291	−1	18	17	93.5
	0.6	0.253	−1	18	17	92.9	0.281	−1	18	17	93.3
600	0.1	0.117	0	7	6	93.6	0.129	0	7	7	93.9
	0.3	0.231	0	12	12	94.5	0.255	0	12	12	94.1
	0.5	0.263	0	14	14	94.5	0.291	0	15	14	94.2
	0.6	0.253	0	14	14	94.1	0.281	0	15	14	93.9

$\hat{R}^1_{adj}(\tau, \hat{\beta})$

n	τ	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)	TRUE	EB ×10³	ESE ×10³	ASE ×10³	C95 (%)
200	0.1	0.478	2	53	51	93.6	0.425	1	54	52	93.9
	0.3	0.473	0	49	47	93.8	0.417	−1	50	48	93.8
	0.5	0.472	−1	51	49	93.4	0.415	−1	52	50	93.7
	0.6	0.473	−1	54	51	93.2	0.416	−1	54	52	93.3
400	0.1	0.478	0	37	36	94.4	0.425	0	39	37	93.7
	0.3	0.473	0	35	33	94.0	0.417	0	35	34	93.4
	0.5	0.472	0	36	35	94.4	0.415	0	37	35	93.4
	0.6	0.473	0	37	36	94.0	0.416	1	39	37	93.4

n	τ	TRUE	EB ×10³	Model A ESE ×10³	ASE ×10³	C95 (%)		TRUE	EB ×10³	Model B ESE ×10³	ASE ×10³	C95 (%)
600	0.1	0.478	−1	30	29	94.5		0.425	−1	30	30	94.7
	0.3	0.473	−1	28	27	94.8		0.417	−1	28	28	94.0
	0.5	0.472	−1	29	28	94.4		0.415	0	29	29	94.3
	0.6	0.473	−1	30	30	94.6		0.416	−1	31	30	94.4

Table 2

Simulation results: empirical rejection rates (ERR) based on $\mathcal{T}_{AB}(\tau)$ and perturbations, where the bolded cells are empirical sizes, and the remaining cells correspond to empirical power. The first row within each section presents the difference in $R^l(\tau)$ between the two models under comparison.

		α = 0.05				α = 0.1			
		τ				τ			
		0.1	0.3	0.5	0.6	0.1	0.3	0.5	0.6
(i) A vs. B	Diff	0.052	0.057	0.057	0.056	0.052	0.057	0.057	0.056
	ERR, $n = 200$	0.611	0.739	0.726	0.675	0.732	0.835	0.822	0.793
	ERR, $n = 400$	0.876	0.959	0.946	0.918	0.930	0.977	0.976	0.963
	ERR, $n = 600$	0.966	0.996	0.997	0.989	0.981	0.997	0.998	0.994
(ii) A vs. C	Diff	0.114	0.021	0.000	0.005	0.114	0.021	0.000	0.005
	ERR, $n = 200$	0.991	0.567	**0.055**	0.134	0.996	0.666	**0.094**	0.212
	ERR, $n = 400$	1.000	0.877	**0.053**	0.259	1.000	0.932	**0.096**	0.350
	ERR, $n = 600$	1.000	0.964	**0.052**	0.349	1.000	0.980	**0.093**	0.475
(iii) E vs. A	Diff	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ERR, $n = 200$	**0.059**	**0.041**	**0.038**	**0.048**	**0.125**	**0.087**	**0.090**	**0.103**
	ERR, $n = 400$	**0.052**	**0.047**	**0.052**	**0.052**	**0.110**	**0.093**	**0.098**	**0.105**
	ERR, $n = 600$	**0.048**	**0.046**	**0.051**	**0.051**	**0.101**	**0.096**	**0.101**	**0.103**
(iv) B vs. C	Diff	0.062	−0.036	−0.057	−0.052	0.062	−0.036	−0.057	−0.052
	ERR, $n = 200$	0.290	0.221	0.570	0.452	0.431	0.333	0.711	0.593
	ERR, $n = 400$	0.527	0.426	0.910	0.767	0.674	0.546	0.948	0.839
	ERR, $n = 600$	0.714	0.574	0.979	0.904	0.818	0.682	0.991	0.945
(v) B vs. D	Diff	0.087	0.015	0.000	0.003	0.087	0.015	0.000	0.003
	ERR, $n = 200$	0.963	0.425	**0.041**	0.108	0.980	0.536	**0.088**	0.174
	ERR, $n = 400$	1.000	0.695	**0.049**	0.179	1.000	0.788	**0.095**	0.255
	ERR, $n = 600$	1.000	0.873	**0.049**	0.263	1.000	0.925	**0.096**	0.375

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Analysis of the dialysis data: list of models and the corresponding $\hat{R}^1$'s.

	Covariates	$\hat{\mathbf{R}}^1$
Model A	AGE, BLACK, FISHH, HCT, LEGS	0.345
Model B	AGE, BLACK, FISHH, HCT	0.327
Model C	AGE, FISHH, HCT, LEGS	0.289
Model D	BLACK, FISHH, HCT, LEGS	0.206
Model E	AGE, BLACK, HCT, LEGS	0.324
Model F	AGE, BLACK, FISHH, LEGS	0.328
Model G	AGE_B, BLACK, FISHH, HCT, LEGS	0.290
Model H	AGE, BLACK, FISHH, HDPD, NEWCAR	0.326
Model I	AGE, BLACK, FISHH, HCT, LEGS, HDPD, HIEDU	0.362

Table 4

Analysis of the dialysis data: results of model comparisons. Each cell corresponds to the p-value at a specific τ or for τ ∈ [0.05, 0.65). The p-values were based on 1,999 perturbed resamples.

Models	Nested	τ							
		0.1	0.2	0.3	0.4	0.5	0.6	[0.05,0.65)
A vs. B	Yes	0.092	0.039	0.4	0.3	0.7	0.5	0.12
A vs. C	Yes	0.028	0.006	0.001	0.004	0.051	0.017	<.001
A vs. D	Yes	0.08	<.001	<.001	<.001	<.001	<.001	<.001
A vs. E	Yes	0.2	0.036	0.034	0.016	0.138	0.3	0.02
A vs. F	Yes	0.04	0.1	0.3	0.1	0.2	0.6	0.05
A vs. G	No	0.2	0.02	0.006	0.01	0.06	0.09	0.02
A vs. H	No	0.3	0.3	0.8	1	0.9	0.9	0.5
B vs. H	No	0.8	0.9	0.8	0.7	0.8	0.7	0.95
I vs. A	Yes	0.3	0.5	0.4	0.3	0.2	0.2	0.2
I vs. B	Yes	0.05	0.08	0.3	0.3	0.2	0.2	0.084