CENSORED QUANTILE REGRESSION SURVIVAL MODELS WITH A CURE PROPORTION

NAVEEN NARISETTY

Department of Statistics, University of Illinois,
725 South Wright Street, Champaign, IL 61820 USA

ROGER KOENKER

Department of Economics, University College London,
London, WC1H 0AX, UK

ABSTRACT. A new quantile regression model for survival data is proposed that permits a positive proportion of subjects to become unsusceptible to recurrence of disease following treatment or based on other observable characteristics. In contrast to prior proposals for quantile regression estimation of censored survival models, we propose a new "data augmentation" approach to estimation. Our approach has computational advantages over earlier approaches proposed by Wu and Yin (2013, 2017). We compare our method with the two estimation strategies proposed by Wu and Yin and demonstrate its advantageous empirical performance in simulations. The methods are also illustrated with data from a Lung Cancer survival study.

Keywords. Survival data, cure proportion, quantile regression, mixture models, data augmentation

1. INTRODUCTION

Motivated to some degree by recent progress in cancer treatment, there has been an increasing interest in survival analysis models that accommodate a probability of "cure," that is a positive treatment effect that lengthens survival prospects to the extent that probability of recurrence or death from the original disease is reduced essentially to zero (Othus et al., 2012). Conventional survival models assume that the survival rate decreases to zero with time going to infinity and cannot be directly used when there is a proportion of subjects getting cured. More flexible survival models for modeling cure rate need to be considered and estimation of such models is obviously challenging since we must distinguish cured subjects from those merely censored by various aspects of the study design and still susceptible to the disease.

In econometrics and the project evaluation literature more generally there are often similar "cure" considerations. For example, in the analysis of unemployment durations, there are often subjects who are never reemployed, some of whom may be interpreted as perpetually cured of the "disease" of work at least in its remunerative forms. See, for example, (Yamaguchi, 1992).

Several statistical models and inference approaches for survival analysis with a cure proportion have been proposed in the literature. There are broadly two classes of commonly used models:

(i) promotion time cure models, which directly model the survival function similar to the Cox-PH model but with the flexibility that the survival function need not go to zero at infinity (Yakovlev and Tsodikov, 1996; Tsodikov, 2002; Bremhorst and Lambert, 2016), and (ii) two component mixture models, where the mixing proportion models the cure rate, and the mixing distributions model the survival functions for the uncured and cured subjects (Kuk and Chen, 1992; Sy and Taylor, 2000; Wu and Yin, 2017; López-Cheda et al., 2017). A comprehensive review of these two approaches along with some methods that unify them is provided in Amico and Van Keilegom (2018).

While both these approaches have their own merits, we consider the mixture model framework as it separates the covariate effects that determine the cure proportion, and the covariate effects that affect the survival time of the uncured subjects (also called latency). The mixture model framework is also the more commonly used one in practice as it allows flexible choices for the survival function of the uncured subjects and for the cure rate proportion. Logistic regression is most commonly used for modeling the cure rate proportion (Kuk and Chen, 1992; Peng and Dear, 2000; Wu and Yin, 2013). There are exceptions such as (Xu and Peng, 2014; López-Cheda et al., 2017) which use nonparametric models for the cure proportion. While parametric or semi-parametric survival functions have long been used (Yamaguchi, 1992; Sy and Taylor, 2000), nonparametric approaches have also been considered in the recent literature (López-Cheda et al., 2017).

Quantile regression (Koenker and Bassett, 1978; Koenker, 2005) provides a more general modeling framework for survival analysis compared to commonly used (semi-) parametric approaches such as Cox PH and AFT models. QR survival models provide a flexible, local specification of covariate effects in the spirit of nonparametric approaches while still maintaining the linear parametric structure familiar from regression modeling. QR survival models and their estimation strategies are studied by Koenker and Geling (2001); Portnoy (2003); Peng and Huang (2008); Wang and Wang (2009); Yang et al. (2018) when there is no cure proportion. More recently, Wu and Yin (2013) proposed a cure rate survival model using quantile regression. Wu and Yin (2013) used a logistic model for the cure proportion and a quantile regression based survival distribution for modeling the latency.

Wu and Yin (2013) initially proposed an estimation strategy that alternated between the estimation of the cure proportion and the latency. However, they acknowledged that the procedure was unstable, and sometimes failed to converge. Wu and Yin (2017) proposed an alternative approach for estimation using multiple imputation (MI). The MI approach first estimates the logistic model using a local Nelson-Aalen type estimator (Wang and Wang, 2009), and imputes the cured subjects followed by applying Portnoy (2003)'s method for estimating the QR model in the latency. While the MI approach has an improved computational performance, it is still limited by the local Nelson-Aalen estimation whose performance deteriorates rapidly as the covariate dimension increases. The linear index specification of quantile regression specification imposes further structure, allowing us to retain the $\sqrt{n}$ convergence rate for the parameters of the survival function for any (fixed) dimension of the covariates, while the local kernel weighting inherits the slower rates associated with nonparametric kernel regression.

We propose a new data augmentation based estimation approach for the cure rate quantile regression model. Our approach provides a more stable estimation algorithm, and is demonstrated through simulation experiments to be more efficient than existing methods especially when there are several predictors. This is to be expected given the rate improvement offered by the global linear index structure of the quantile regression model. Our method is motivated by recent work on using data augmentation for censored quantile regression (Yang et al., 2018), but significantly

generalizes this approach for dealing with cure proportion. More specifically, our method augments both the cure indicators as well as the censored responses and iteratively updates the quantile regression coefficients, cure rate parameters and the augmented variables. Each step of the update only involves convex functions making it computationally efficient. A distinct advantage of our approach compared to existing alternatives is that it can more efficiently incorporate multiple covariates as required in most applications.

We now provide an outline of the paper. In Section 2 we introduce the cure rate quantile regression model and existing methods for estimation. In Section 3, we describe our proposed estimation method. In Section 4, we discuss an implementation of our method in R. We provide simulation results and application of our method to a lung cancer study in Section 6 followed by a discussion in Section 7.

2. CURE RATE QR MODEL

The most basic quantile regression survival model as introduced in Koenker and Geling (2001), assumes that the τth conditional quantile functions of the possibly transformed survival time T are given by,

$$Q_{T_i}(\tau|X = x_i) = x_i^\top \beta(\tau). \tag{1}$$

Portnoy (2003) and Peng and Huang (2008) proposed estimation methods for this model that accounted for the almost inevitable presence of censoring. Portnoy (2003) built upon an analogy with the well-known univariate Kaplan-Meier estimator, while Peng and Huang (2008) built upon the martingale representation afforded by the univariate Nelson-Aalen estimator.

An advantage of the QR survival model is that it allows the researcher to be quite flexible about how the covariates enter into the model locally at each quantile level of the response, while maintaining the linear parametric structure familiar from regression modeling. From an asymptotic viewpoint this is reflected in parametric rates of convergence for the estimator of $\beta(\tau)$. The downside of this in the presence of censoring is that it requires a global (linear) specification of the covariates effects in order to justify the weighting schemes used to account for the censoring (Portnoy, 2003). Nevertheless, the global quantile regression model, a model satisfying Equation (1) at all quantile levels, represents a large class of models encompassing heteroskedastic models such as location-scale models with covariate-dependent location and scale. The classical survival models such as the proportional hazards (PH) model and the accelerated failure time (AFT) model are special cases of the global QR model with a transformed survival time as the response and the slope coefficients of $\beta(\tau)$ constant across τ.

When there is a cure proportion, the possibility of a cure is introduced via a latent variable, η, modeled as a binary response. The probability of subject i being susceptible (not cured), denoted by π_i, depends on covariates Z as mediated by the link function, π. That is,

$$\pi_i = \mathcal{P}(\eta_i = 1|Z = z_i) = \pi(z_i^\top \gamma), \tag{2}$$

As in Wu and Yin, we use the logistic link $\pi(u) = e^u/(1 + e^u)$, but it is possible to consider other potential choices such as the Gosset link functions (Koenker and Yoon, 2009) or fully nonparametric link functions (Xu and Peng, 2014; López-Cheda et al., 2017). When $\eta_i = 1$ we will say that subject i is susceptible to the event of interest, while if $\eta_i = 0$ they are unsusceptible, thus,

$$\tilde{Y} = \eta T + (1 - \eta)\infty, \tag{3}$$

subject to the usual constraints of censoring. We observe, $Y_i = \tilde{Y}_i \wedge C_i$, where C_i denotes a random censoring time, and $\Delta_i = I(\tilde{Y}_i \leqslant C_i)$. We will assume, further, that Y and C are conditionally independent given the covariates X and Z. Under these conditions, our objective is to estimate the cure rate parameters γ and the QR parameters $\beta(\tau)$.

2.1. **Existing Estimation Methods for Cure Rate QR Model.** For censored quantile survival model, Peng and Huang (2008) use a martingale based on the counting process $N_i(t) = \Delta_i I(Y_i < t)$ to construct an estimating equation for the quantile regression process. A comprehensive treatment of survival analysis from this viewpoint is available from Anderson et al. (1993). Wu and Yin (2013) generalized this approach to cure rate quantile model. More specifically, using the cumulative hazard function,

$$\Lambda_Y(t|x_i, z_i) = -\log(1 - \pi(z_i^\top \gamma) F_T(t|x_i)),$$

where $F_T(t|x_i)$ is the conditional distribution of T given x_i, we have the martingale,

$$M_i(t) = N_i(t) - \Lambda_Y(t|x_i, z_i),$$

with respect to the natural filtration of information up to time t. This standard counting process formulation of the Nelson-Aalen estimator can be employed to construct an estimating equation for γ given an estimator for F_T. Building on the prior work of Beran (1981), Dabrowska (1987) and others, Wang and Wang (2009) proposed estimating censored QR models using a local, kernel weighted version of the Kaplan-Meier estimator for F_T. Wu and Yin (2013) adopt this approach and construct a locally weighted Nelson-Aalen estimator. For cure applications, this has the advantage that an estimating equation for γ can be constructed that avoids any global parametric specification of the quantile specific effects. The difficulty with their approach, of course, is that specification of the kernel and associated bandwidths becomes increasingly problematic as the dimension of the covariate space grows. Given an estimator for γ, Wu and Yin (2013) construct another set of estimating equations for $\beta(\tau)$ in the same spirit as Peng and Huang (2008). While an iterative procedure to estimate $\beta(\tau)$ and γ alternatively is proposed by Wu and Yin (2013), they note convergence issues of this approach due to the complexity of the iterating steps.

Wu and Yin (2017) extend their prior approach by noting that the conditional probabilities of subjects being susceptible can be computed from the estimator of γ obtained by the local Nelson-Aalen method and used to impute η's for the full sample. Of course, for subjects with $\Delta_i = 1$ these probabilities are necessarily one as they correspond to susceptible subjects. Once the η's are imputed, the corresponding susceptible subjects are used to estimate $\beta(\tau)$ similar to Wang and Wang (2009). This imputation process is performed until some criterion of convergence is achieved. Such imputation schemes can be expected to improve upon the earlier estimating equation method, but it still suffers from the inherent drawbacks of the local Nelson-Aalen approach.

3. Proposed Estimation Method using Data Augmentation

We now describe our proposed approach for estimating γ and $\beta(\tau)$ based on data augmentation. Our data augmentation estimator generalizes the approach of Yang et al. (2018) to cure rate quantile model and shares some features of the imputation method. In addition to augmenting the censored observations, we augment the latent indicators η_i's for deriving a data augmentation-like algorithm. In contrast to the existing methods which rely on local nonparametric methods, however, our data augmentation relies only on the global parametric specification of the QR process allowing us to more easily accommodate several covariates in X.

Our data augmentation starts with initial values for the QR process, $\beta(\tau)$ on the grid $\tau_1, \ldots, \tau_M$ which can be obtained by simply computing the median regression estimator $\hat{\beta}(1/2)$, based on only the uncensored observations and imposing the common slope assumption, so $\hat{\beta}(\tau) = \hat{\beta}(1/2) + \hat{\beta}_1(\tau)e_1$ where $\hat{\beta}_1(\tau)$ denotes the ordinary sample quantiles of the residuals from the median fit and e_1 is the first unit basis vector of $\mathbb{R}^p$. An initial estimator of γ is obtained by (naively) estimating the binary response model of δ on Z, i.e. assuming provisionally that all the censored subjects are cured. Although data augmentation does not demand a consistent initialization, estimators from existing methods can also be used for initialization to achieve faster convergence. Given these initial estimators, we may begin the iteration by performing each of the following steps conditional on all the remaining quantities:

- Generate η_i's,
- Reestimate γ,
- Generate the censored y_i's,
- Reestimate $\beta(\tau)$.

Accumulating the $\hat{\gamma}$'s and $\hat{\beta}(\tau)$'s from this iteration, point estimates can be obtained by simply averaging over the corresponding iterates. For both reestimation steps, there is the option to resample with replacement from the relevant full sample as in the standard (x, y) bootstrap. We now provide a more comprehensive description of the data augmentation approach in the following algorithmic structure.

Step 0 (Initialization): Initialize $\hat{\boldsymbol{\beta}}^{(0)}(\tau_k)$ for $k = 1, \cdots, M = \max\{\lfloor\sqrt{n}\rfloor, 100\}$ and $\hat{\gamma}^{(0)}$.

Step 1 (Data Augmentation): Given the estimates $\hat{\boldsymbol{\beta}}^{(h)}(\tau_k)$ and $\hat{\gamma}^{(h)}$, perform the following sampling steps at iteration $(h+1)$ in the order of their appearance:

- Generate $\eta_i^{(h+1)}$'s based on the conditional distribution $(\eta \mid \Delta, X, Z, Y)$ using the current estimates $\hat{\boldsymbol{\beta}}^{(h)}$ and $\gamma^{(h)}$ of the parameters. That is, generate $\eta_i^{(h+1)}$ as Bernoulli draw with probability given by

$$\hat{\pi}_i := P[\eta = 1 \mid \Delta, X, Z, Y] = \Delta + (1-\Delta)\frac{\pi(z^\top\gamma^{(h)})(1-\hat{F}_T(Y \mid X))}{1-\pi(z^\top\gamma^{(h)})\hat{F}_T(Y \mid X))},$$

  where $\hat{F}_T(Y \mid X)$ is the estimated CDF of $T \mid X$ corresponding to the regression quantiles $\hat{\boldsymbol{\beta}}^{(h)}$ evaluated at the observed Y.
- Sample γ from the bootstrap (posterior) distribution of γ given the data $\{\eta_i^{(h+1)}, z_i\}, i = 1, \cdots, n$. For sampling from the bootstrap distribution, obtain a resample of size (equal probability with replacement) from the data $\{\eta_i^{(h+1)}, z_i\}, i = 1, \cdots, n$ and set $\hat{\gamma}^{(h+1)}$ as the corresponding MLE.

$$\hat{\gamma}^{(h+1)}(\tau) \leftarrow \arg\max_{\gamma} L(\gamma \mid \{\eta_i^{(h+1)}, z_i\}),$$

  where $L(\cdot)$ is the logistic likelihood given by

$$L(\gamma \mid \{\eta_i^{(h+1)}, z_i\}) \propto \prod_{i=1}^{n} \frac{\exp\{\eta_i^{(h+1)} z_i^\top \gamma\}}{\left(1+\exp\{z_i^\top\gamma\}\right)}.$$

Alternatively, if a prior on γ is available, $\hat{\gamma}^{(h+1)}$ is sampled from the posterior distribution:

$$\pi\left(\gamma \mid \{\eta_i^{(h+1)}, z_i\}\right) \propto L(\gamma \mid \{\eta_i^{(h+1)}, z_i\}) \times \pi(\gamma),$$

where $\pi(\gamma)$ is the prior distribution on γ.

- Generate the censored y_i's from their conditional distribution given the estimates $\hat{\beta}^{(h)}(\tau_k)$ and $\eta_i^{(h+1)}$'s. That is, if $\delta_i = 0$ and $\eta_i^{(h+1)} = 0$, so y_i takes the value infinity as these correspond to the cured subjects. To generate y_i's when $\delta_i = 0$ and $\eta_i^{(h+1)} = 1$, define k_m to be the first index such that $x_i^T \hat{\beta}^h(\tau_m) \geqslant C_i$ (provided such an m exists). Draw a random number τ_i^* uniformly from $\{\tau_k : k = m, \cdots, M\}$ and set $y_i^{(h)} = x_i^T \hat{\beta}^{(h)}(\tau^*)$.
- Sample $\beta(\tau)$: collect observations with $\eta_i^{(h+1)} = 1$, and sample $\beta_i^{(h+1)}$ from the bootstrap (posterior) distribution of β given the data $(y_i^{(h)}, x_i), i : \eta_i^{(h+1)} = 1$. That is, obtain a resampled data of the same size from the uncured observations, and estimate $\hat{\beta}^{(h+1)}(\tau)$ from the usual quantile regression estimator:

$$\hat{\beta}^{(h+1)}(\tau) \leftarrow \arg\min_{\beta} \sum_{i:\, \eta_i^{(h+1)}=1} \rho_\tau(y_i^{(h)} - x_i^T \beta),$$

where $\rho_\tau(u) = \tau|u| - u\mathbb{1}\{u < 0\}$ is the check loss function. For the ease of notation, we suppressed additional notation required to indicate the bootstrap sample is to be used in the optimization above.

Step 2 (Aggregation): Iterate Step 1 for a pre-specified number of iterations H or until a specified convergence criterion is met. The final estimate $\tilde{\beta}(\tau)$ and $\tilde{\gamma}$ are obtained by averaging the estimates from the last half of the iterations. That is,

$$\tilde{\beta}(\tau) \leftarrow \sum_{h=1}^{H} \hat{\beta}^{(h)}(\tau), \text{ and } \tilde{\gamma} \leftarrow \sum_{h=1}^{H} \hat{\gamma}^{(h)}.$$

We note here that the sampling of $\beta(\tau)$ in Step 1 could be performed using a posterior distribution in place of bootstrap distribution. However, this would require the use of a working likelihood as the true likelihood is difficult to deal with along with a prior specification on all the regression quantiles $\beta(\tau_k)$ for $k = 1, \cdots, M$. While this is certainly possible, we have chosen to rely on the bootstrap sampler for its simplicity.

We now offer a heuristic, Bayesian interpretation of our final estimators $\tilde{\beta}(\tau)$ and $\tilde{\gamma}$. The distribution of Y, Δ, C given $\{\beta(\tau), \tau \in (0,1)\}, \eta, X, Z$, can be interpreted as a likelihood function for estimating $\beta(\tau)$ and η. The obvious practical difficulty is that we do not have a simple closed form expression for the likelihood suitable for optimization or simulation. Instead, we consider an augmented likelihood with η serving as the augmented data, i.e., we consider the distribution of Y, η, Δ, C given $\{\beta(\tau), \tau \in (0,1)\}, \gamma, X, Z$ and denote the corresponding likelihood as $L(\beta, \gamma \mid \{Y, \Delta, \eta, C, X, Z\})$. If we had prior distributions $\pi(\beta)$ and $\pi(\gamma)$ on the parameters, β, and γ then the corresponding posterior distribution would take the form: $\pi(\beta, \gamma \mid \{Y, \Delta, C, X, Z\}) \propto L(\beta, \gamma \mid \{Y, \Delta, \eta, C, X, Z\})\pi(\beta)\pi(\gamma)$. While this posterior distribution is available in principle, it is still somewhat intractable. If we proceed in a Gibbs sampling like conditional sampling mode, the conditional sampling steps for γ and η are relatively straightforward as provided in our data augmentation algorithm. The computationally challenging aspect is the conditional distribution of

β since this is the difficult component of the posterior distribution. To circumvent this issue, our proposed data augmentation algorithm estimates regression quantiles on a discrete grid of quantile levels and then makes use of the bootstrap distribution as an approximation to the intractable conditional distribution of β. The resultant data augmentation estimators can be viewed as approximate posterior mean estimators for β and γ where the approximation is due to the quantile discretization as well as the bootstrap approximation of the conditional distribution of β. While quantile discretization is commonly employed in quantile regression inference (Koenker, 2005), bootstrap approximation to the conditional distribution of $\hat{\beta}(\tau)$ in simulation settings with latent variables is more recent, notably in Yang et al. (2018) for estimation in censored quantile regression, and Arellano et al. (2017) for dynamic panel models of income dynamics.

4. Software Implementation

In this section, we will briefly describe the R implementation of the foregoing methods. We have developed an R function that provides a unified interface to all the three estimation methods. The function is `cqr()`, pronounced "cure," not to be confused with `crq()`, which is the umbrella function for censored quantile regression applications in the R package **quantreg**. We expect eventually to try to fold the functionality of `cqr` into **quantreg** and perhaps even into `crq`, but for the moment it seems prudent to keep them separate.

The `cqr` function uses the extended formula interface of the package **Formula**, so one writes the model as `y | d ~ X | Z` where `y` denotes the observed response, `d` the censoring indicator, `X` the covariates of the QR model, and `Z` the covariates of the binary response model. The remaining arguments are standard, with the `method` argument taking one of three possible values, `LNA`, `Imp` or `DA` corresponding to the three methods:

- `LNA`: Local Nelson Aalen estimation method of Wu and Yin (2013),
- `Imp`: Imputation method of Wu and Yin (2017), and
- `DA`: Our proposed data augmentation method.

The `DA` method by default does not use a bootstrap resample at the sampling step of $\beta(\tau)$. To implement a bootstrapped version, the option `bootstrap = TRUE` can be provided which we will denote by `DA.B` in the remaining part of the paper. Users have the option of specifying a vector of τ's of interest when evaluating $\hat{\beta}(\tau)$ as well as the grid of τ's used for the intermediate computations. The latter, by default, is set to the percentiles.

The default link function for the binary response cure component of the model is logistic, but other link functions compatible with the R `glm` function are easily available. These include probit and cauchit, but one could also use any of the parametric links available from the package **glmx**, Zeileis et al. (2015).

5. Identifiability Considerations

As noted by Patilea and van Keilegom (2017) cure rate models have delicate identifiability requirements. This should not be surprising since we are claiming to distinguish heavy tail behavior of the survival distribution from circumstances in which the event probability is actually zero. We address such considerations in this section. We can rewrite the data generating process as

$$\tilde{Y} = \eta T + (1-\eta)\infty,\ Y = \eta(T \wedge C) + (1-\eta)C.$$

Now, let us suppose that the distribution of the censored time C has support on $[0, M)$, for some $0 < M \leqslant \infty$. The following theorem shows that the model is identifiable if $M = \infty$ but not necessarily identifiable if $M < \infty$. The censoring distribution must offer some hope of observing the entire tail of the uncured survival distribution. Recall that our observed data consists of $Y = (\tilde{Y} \wedge C)$ and $\Delta = 1\{\tilde{Y} \leqslant C\}$.

Theorem 1. *Suppose that*
(i) the parameter γ is identifiable given observable η. That is, if $\pi(z^\top \gamma_1) = \pi(z^\top \gamma_2)$ for all z, then $\gamma_1 = \gamma_2$, where recall that $\pi(z^\top \gamma) = P_\gamma[\eta = 1 \mid Z = z]$: and
(ii) the regression quantiles $\beta(\tau)$ are identifiable given observable T. That is, if $P_{\beta_1}[T \leqslant u \mid X = x] = P_{\beta_2}[T \leqslant u \mid X = x]$ for all $u \in (0, \infty)$ and for all x implies that $\beta_1 = \beta_2$. Then:

- *If $M = \infty$, then all the parameters of the model are identifiable.*
- *If $M < \infty$, then the parameters of the model may not necessarily be identifiable.*

Proof. Let us consider the conditional distribution of $(T \wedge C)$ which is given by:

$$\begin{aligned} P[(T \wedge C) \leqslant u \mid X = x, Z = z] &= 1 - P[T > u \mid X = x]P[C > u \mid X = x] \\ &= 1 - (1 - F_T(u \mid x))(1 - F_C(u \mid x)). \end{aligned} \tag{4}$$

If for two sets of parameters (γ_1, β_1) and (γ_2, β_2), the corresponding conditional distributions of $\tilde{Y}$ are the same, then the distributions for $P[Y \mid X = x, Z = z]$ and $P[\Delta = 1 \mid X = x, Z = z]$ based on the two parameter sets need to be identical. Consider:

$$\begin{aligned} P[\Delta = 1 \mid X = x, Z = z] &= P[\tilde{Y} \leqslant C \mid X = x, Z = z] \\ &= P[\eta = 1 \mid Z = z]P[T \leqslant C \mid X = x] \\ &= P_\gamma[\eta = 1 \mid z] \int_0^M P_\beta[T \leqslant u \mid x] dF_C(u|x), \end{aligned} \tag{5}$$

If this distribution is the same for two different combinations of the parameters for all possible distributions of the censoring time C, we will need that: $P_{\gamma_1}[\eta = 1 \mid Z = z]P_{\beta_1}[T \leqslant u \mid X = x] = P_{\gamma_2}[\eta = 1 \mid Z = z]P_{\beta_2}[T \leqslant u \mid X = x]$, for almost all $u \in (0, M)$. If $M = \infty$, by letting u tend to ∞, we obtain that $P_{\gamma_1}[\eta = 1 \mid z] = P_{\gamma_2}[\eta = 1 \mid z]$. Since this holds for all z, assumption (i) on the identifiability of γ necessarily implies that $\gamma_1 = \gamma_2$.

Now, consider the distribution of $Y \mid X, Z$.

$$P_{\beta,\gamma}[Y \leqslant u \mid X = x, Z = z] = 1 - (1 - P_\gamma[\eta = 1 \mid Z = z]P_\beta[T \leqslant u \mid X = x](1 - F_C(u \mid x)) \tag{6}$$

If the above distributions are to be the same for two sets of parameters (γ_1, β_1) and (γ_2, β_2), in the case $M = \infty$, we have $\gamma_1 = \gamma_2$ using the previous argument. Therefore, since F_C does not depend on β and γ, it immediately follows that $P_{\beta_1}[T \leqslant u \mid X = x] = P_{\beta_2}[T \leqslant u \mid X = x]$ for all $u \in (0, \infty)$. This implies that $\beta_1 = \beta_2$ due to condition (ii) of the theorem.

Finally, consider the joint distribution of Y and Δ when $M < \infty$. For $y \in (0, M)$, we have

$$\begin{aligned}
P[Y \leqslant y, \Delta = 1 \mid X = x, Z = z] &= P[(\tilde{Y} \wedge C) \leqslant y, (\tilde{Y} \leqslant C) \mid X = x, Z = z] \\
&= \int_0^M P[\tilde{Y} \leqslant y, (\tilde{Y} \leqslant C) \mid X = x, Z = z] dF_C(u|x) \\
&= \int_0^y P[\tilde{Y} \leqslant u \mid X = x, Z = z] dF_C(u|x) + P[\tilde{Y} \leqslant y \mid X = x, Z = z] P[y \leqslant C \mid X = x] \\
&= P[\eta = 1 \mid Z = z] \left(\int_0^y P[T \leqslant u \mid X = x] dF_C(u|x) + P[T \leqslant y \mid X = x] P[y \leqslant C \mid X = x] \right).
\end{aligned}$$

Therefore, from the calculations in Equations (6), for any pair of parameter values, the joint distribution of (Y, Δ) does not change if and only if for all $u \in (0, M)$, we have

$$P_{\gamma_1}[\eta = 1 \mid Z = z] P_{\beta_1}[T \leqslant u \mid X = x] = P_{\gamma_2}[\eta = 1 \mid Z = z] P_{\beta_2}[T \leqslant u \mid X = x]. \tag{7}$$

For this to be satisfied, it is not necessary that $\gamma_1 = \gamma_2$ and $\beta_1 = \beta_2$. To see this, consider an example with a single binary covariate W so that $X = Z = (1, W)$. For any specific choice of $P_{\beta_1}[T \leqslant u \mid (W = 0)]$, define β_2 so that

$$P_{\beta_2}[T \leqslant u \mid (W = 0)] = \frac{P_{\gamma_2}[\eta = 1 \mid (W = 0)]}{P_{\gamma_1}[\eta = 1 \mid (W = 0)]} P_{\beta_1}[T \leqslant u \mid (W = 0)], \text{ for } u \in (0, M).$$

Due to the identifiability condition (i), the link function $\pi(\cdot)$ is non-constant and hence the intercepts of γ_1 and γ_2 can be chosen such that

$$\frac{P_{\gamma_2}[\eta = 1 \mid (W = 0)]}{P_{\gamma_1}[\eta = 1 \mid (W = 0)]} < 1.$$

With such choices for γ_1 and γ_2, the intercept process of β_2 can be defined as

$$\beta_2^0(\tau) = \beta_1^0 \left(\frac{P_{\gamma_2}[\eta = 1 \mid (W = 0)]}{P_{\gamma_1}[\eta = 1 \mid (W = 0)]} \tau \right) = \Phi^{-1}(\tau).$$

Similarly, we can define the slope for all the parameters based on the conditional distribution at $W = 1$, which proves non-identifiability of the parameters if $M < \infty$. □

The proof indicates that identifiability of the model depends crucially on the censoring distribution which is to be expected. In the context of clinical trials, this implies that if the duration of the study is relatively a short, one needs to worry about identifiability considerations quite seriously. While the theorem suggests that the model is not necessarily identifiable, it does not automatically imply non-identifiability for every design. The counter-example in the proof uses a binary predictor, but if the design is well-chosen and the predictor space is sufficiently rich identifiability issues may not arise. However, to characterize the identifiability explicitly for a given design is rather difficult due to the analytic intractability of its interaction with the form of the regression quantile process. However, one might numerically check the validity of the identifiability condition by checking whether Equation (7) implies $(\gamma_1, \beta_1) = (\gamma_2, \beta_2)$. Thus the proof of the theorem also provides a broader characterization of identifiability of the cure rate quantile regression model.

	p	γ	β	L	Censoring rate	Cure rate
Case 1	1	$(1,-1)$	$(2,1)$	40	0.41	0.38
Case 2	1	$(-0.5,1)$	$(1,-1)$	4	0.56	0.50
Case 3	9	$(1,-1,1,\cdots,-1)$	$-\gamma$	4	0.41	0.39
Case 4	1	$(0.25,1)$	$(2,1)$	4	0.60	0.32
Case 5	1	$(0.25,1)$	$(2,1)$	6	0.51	0.32

TABLE 1. Different cases considered in our simulation studies

6. EMPIRICAL STUDIES

6.1. **Simulations.** We first consider the following simulation set-up of Wu and Yin (2017):

$$\log(\pi_i/(1-\pi_i)) = \gamma_0 + \gamma_1 x_i,$$

and the event time model,

$$Y_i = \log T_i = \beta_0 + \beta_1 x_i + (1+x_i)u_i$$

where $u \sim \mathcal{N}(0,1)$. Censoring is determined by x and a random uniform, $R \sim U[0,L)$ as,

$$C_i = I(x_i < 1/2)R_i + I(x_i \geqslant 1/2)(R_i+1).$$

Wu and Yin (2017) set $\gamma = (1,-1)$, $\beta = (2,1)$ and $L = 40$. Note, however, that this L, which represents the duration of the study in clinical trial applications corresponds to a rather unrealistic, essentially infinite value. Therefore, we expand the simulation settings to other choices of γ, β, L, and p totalling five different cases reported in Table 1.

Case 1 is the same as the setting of Wu and Yin (2017). Cases 2-5 have much smaller value for L (either 4 or 6) so that the study duration is more realistic. Case 2 has large censoring and cure rates. Case 3 has $p = 9$ covariates representing a multiple regression scenario. Cases 4 and 5 have a high censoring rate and a moderate cure rate. In the last two cases, the initial estimator for γ is heavily biased and unreliable in contrast to the first three cases. Moreover, the high censoring rate in Case 4 makes the upper conditional quantiles of the latency unidentifiable, since subjects with extremely high survival times in the uncured subpopulation cannot be distinguished from the cured subjects.

We report results for the four methods mentioned in Section 4 in terms of both bias and mean squared error (MSE) in Tables 2 - 6 for both sample sizes $n = 200$ and $n = 500$. The experiment is based on 1000 replications.

Our empirical findings can be briefly summarized as follows:

(i) Overall, our proposed data augmentation approaches (DA and DA.B) have lower MSE values when compared to the existing approaches of Wu and Yin (2013, 2017) for estimation of the logistic parameter γ and the quantile regression parameters.

(ii) In some cases (for e.g. Cases 1 and 2), data augmentation based methods have larger bias but they still have smaller MSE indicating that their reduced variability compensates for the bias.

(iii) When there are several predictors (Case 3 with $p = 9$), the performance of LNA and Imp suffer much more than DA. This is expected because those approaches rely on the local Nelson-Aalen method, which is unreliable when the dimension of the covariate space is large.

	Bias				MSE			
	LNA	Imp	DA	DA.B	LNA	Imp	DA	DA.B
				$n = 200$				
γ								
Intercept	0.036	0.036	0.010	0.025	0.106	0.106	0.104	0.114
Slope	−0.026	−0.026	−0.015	−0.032	0.303	0.303	0.301	0.325
β**(0.5)**								
Intercept	0.032	−0.003	−0.076	−0.075	0.076	0.073	0.077	0.077
Slope	0.044	0.010	−0.029	−0.030	0.353	0.348	0.321	0.320
β**(0.7)**								
Intercept	0.056	0.009	−0.055	−0.055	0.088	0.081	0.080	0.080
Slope	0.033	−0.013	−0.072	−0.072	0.408	0.390	0.359	0.360
				$n = 500$				
γ								
Intercept	0.022	0.022	0.007	0.012	0.041	0.041	0.040	0.041
Slope	−0.017	−0.017	−0.010	−0.015	0.118	0.118	0.117	0.118
β**(0.5)**								
Intercept	0.020	−0.003	−0.073	−0.073	0.029	0.029	0.034	0.034
Slope	0.040	0.013	−0.022	−0.022	0.141	0.141	0.135	0.135
β**(0.7)**								
Intercept	0.030	−0.002	−0.060	−0.060	0.033	0.033	0.034	0.035
Slope	0.043	0.004	−0.041	−0.041	0.160	0.155	0.146	0.146

TABLE 2. Case 1: p = 1, censoring rate = 0.41, cure rate = 0.38; Bias and Mean Squared Error for different estimation methods based on 1000 replications

(iv) The performance of all methods is poor for Case 4 due to high bias. The bias in γ suggests that many uncured subjects are classified as cured. This would naturally cause bias in estimation of β as well. As mentioned earlier, due to the high censoring rate, the latency distribution is not fully observed which violates the identifiability condition discussed in Patilea and van Keilegom (2017). This underscores the need to be cautious using cure rate models with high censoring involved in latency that could result in overly optimistic assessments about cure rate proportion. Since DA uses the information in the latency distribution more efficiently, we can see that the impact of this on DA is relatively less compared to LNA and Imp.

6.2. **Lung Cancer Study.** Finally, we briefly reconsider the lung cancer study considered in Wu and Yin (2017), employing the same model as Wu and Yin. We report results from all three fitting methods. The data consists of 280 observations with 64% censoring. There are three covariates: tumor histology, patient age and patient gender. All three are used in both the logistic cure model and the QR survival model. Although we have used the same bandwidth parameters for the local Nelson-Aalen estimation for the "LNA" and "Imp" estimators, our estimates differ slightly from those reported in Wu and Yin (2017). Table 7 reports γ estimates for the three methods, while Figure 1 depicts $\beta(\tau)$ estimates. Standard errors and pointwise confidence bands are based on 200 replications.

	Bias				MSE			
	LNA	Imp	DA	DA.B	LNA	Imp	DA	DA.B
				$n = 200$			]	
γ								
Intercept	0.057	0.057	−0.060	−0.065	0.116	0.116	0.109	0.116
Slope	−0.036	−0.036	−0.011	0.002	0.321	0.321	0.304	0.322
β**(0.5)**								
Intercept	0.079	0.024	−0.239	−0.237	0.138	0.127	0.144	0.143
Slope	−0.013	−0.017	0.199	0.198	0.465	0.452	0.420	0.422
β**(0.7)**								
Intercept	0.142	0.020	−0.313	−0.312	0.206	0.150	0.199	0.198
Slope	−0.079	−0.048	0.165	0.164	0.608	0.495	0.399	0.399
				$n = 500$			]	
γ								
Intercept	0.063	0.063	−0.054	−0.058	0.049	0.049	0.045	0.046
Slope	−0.044	−0.044	−0.011	−0.004	0.126	0.126	0.118	0.119
β**(0.5)**								
Intercept	0.073	0.022	−0.240	−0.239	0.062	0.055	0.093	0.093
Slope	−0.012	0.004	0.225	0.226	0.192	0.192	0.208	0.208
β**(0.7)**								
Intercept	0.112	0.017	−0.307	−0.306	0.082	0.062	0.136	0.136
Slope	−0.027	0.002	0.217	0.217	0.233	0.207	0.201	0.201

TABLE 3. Case 2: p = 1, censoring rate = 0.56, cure rate = 0.5; Bias and Mean Squared Error for different estimation methods based on 1000 replications

Again we see that the three methods produce similar conclusions. In our judgment, the data augmentation approach is preferable for several reasons. It is less sensitive to the upper tail of quantile regression model, it is more easily adaptable to several covariates, and it avoids inherently delicate bandwidth selection issues.

7. DISCUSSION

Quantile regression methods offer an attractive approach to estimating survival models with a positive cure proportion. Covariate effects are flexibly modeled in the upper tail where the cure effect is most salient. Here, we have adopted the modeling strategy of Wu and Yin (2013) and Wu and Yin (2017), however their estimation methods, which are based on the local Nelson-Aalen approach of Wang and Wang (2009) are compared with an alternative data augmentation approach proposed recently by Yang et al. (2018). The latter approach has a number of advantages, and it is the approach we would recommend for most applications.

While we use a logistic regression model for the cure proportion, alternative nonparametric approaches proposed in the recent literature (Wang et al., 2012; Xu and Peng, 2014; Koenker and Yoon, 2009; López-Cheda et al., 2017) can also be used.

	Bias				MSE			
	LNA	Imp	DA	DA.B	LNA	Imp	DA	DA.B
				$n = 200$				
γ								
Intercept	0.664	0.664	−0.008	0.070	1.725	1.725	0.831	0.975
Slope	−0.028	−0.028	0.005	−0.004	0.628	0.628	0.403	0.493
β**(0.5)**								
Intercept	0.240	0.190	−0.020	−0.018	0.383	0.334	0.261	0.261
Slope	0.001	0.000	0.000	0.000	0.157	0.136	0.105	0.105
β**(0.7)**								
Intercept	0.500	0.257	−0.053	−0.052	2.716	0.511	0.271	0.271
Slope	0.008	0.005	0.000	0.000	0.793	0.193	0.119	0.119
				$n = 500$				
γ								
Intercept	0.653	0.653	0.010	0.033	0.983	0.983	0.334	0.358
Slope	−0.030	−0.030	−0.001	−0.004	0.235	0.235	0.141	0.153
β**(0.5)**								
Intercept	0.210	0.194	−0.029	−0.029	0.155	0.140	0.083	0.083
Slope	0.006	0.006	0.003	0.003	0.065	0.058	0.036	0.036
β**(0.7)**								
Intercept	0.429	0.309	−0.051	−0.051	0.449	0.257	0.103	0.103
Slope	0.008	0.008	0.004	0.004	0.179	0.093	0.042	0.043

TABLE 4. Case 3: p = 9, censoring rate = 0.41, cure rate = 0.4; Bias and Mean Squared Error for different estimation methods based on 1000 replications

8. ACKNOWLEDGEMENTS

The initial phase of this research was partially supported by Bristol-Myers Squibb. Both authors would like to acknowledge valuable conversations with Xuming He about the subject matter addressed. The second author would like to express his appreciation to Gary Chamberlain for his encouragement and inspiration over several decades, beginning with his two early JRSS(B) papers, (Chamberlain and Leamer, 1976; Leamer and Chamberlain, 1976), that introduced data augmentation in a remarkably general setting to econometrics.

REFERENCES

Amico, M. and Van Keilegom, I. (2018). Cure models in survival analysis. *Annual Reviews of Statistics and its Applications* **5,** 311–342.

Anderson, P., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Verlag.

Arellano, M., Blundell, R., and Bonhomme, S. (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica* **85,** 693–734.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report: University of California, Berkeley.

	Bias				MSE			
	LNA	Imp	DA	DA.B	LNA	Imp	DA	DA.B
				$n = 200$				
γ								
Intercept	0.144	0.144	−0.168	−0.164	0.192	0.192	0.174	0.184
Slope	−1.110	−1.110	−0.982	−0.980	1.694	1.694	1.365	1.390
β**(0.5)**								
Intercept	0.167	0.089	−0.506	−0.505	0.143	0.113	0.316	0.315
Slope	−0.790	−0.821	−0.619	−0.618	1.018	1.039	0.591	0.589
β**(0.7)**								
Intercept	0.238	0.117	−0.431	−0.432	0.191	0.127	0.260	0.260
Slope	−1.190	−1.211	−1.090	−1.086	1.807	1.827	1.435	1.430
				$n = 500$				
γ								
Intercept	0.142	0.142	−0.164	−0.163	0.085	0.085	0.076	0.077
Slope	−1.116	−1.116	−0.987	−0.986	1.421	1.421	1.127	1.128
β**(0.5)**								
Intercept	0.127	0.066	−0.510	−0.510	0.062	0.048	0.287	0.288
Slope	−0.765	−0.797	−0.616	−0.614	0.733	0.784	0.470	0.468
β**(0.7)**								
Intercept	0.185	0.074	−0.447	−0.447	0.087	0.053	0.232	0.232
Slope	−1.146	−1.160	−1.067	−1.065	1.459	1.494	1.243	1.240

TABLE 5. Case 4: p = 1, censoring rate = 0.6, cure rate = 0.32; Bias and Mean Squared Error for different estimation methods based on 1000 replications

Bremhorst, V. and Lambert, P. (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Comput. Statist. Data Anal.* **93,** 270–284.

Chamberlain, G. and Leamer, E. E. (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society. Series B (Methodological)* **38,** 73–84.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* **14,** 181–197.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association* **96,** 458–468.

Koenker, R. and Yoon, J. (2009). Parametric links for binary choice models: A Fisherian-Bayesian colloquy. *Journal of Econometrics* **152,** 120–130.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79,** 531–541.

Leamer, E. E. and Chamberlain, G. (1976). A Bayesian interpretation of pretesting. *Journal of the Royal Statistical Society. Series B (Methodological)* **38,** 85–94.

López-Cheda, A., Caoa, R., Jácomea, A., and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics*

	Bias				MSE			
	LNA	Imp	DA	DA.B	LNA	Imp	DA	DA.B
				$\mathfrak{n} = 200$				
γ								
Intercept	0.187	0.187	−0.062	−0.057	0.191	0.191	0.136	0.143
Slope	−0.274	−0.274	−0.293	−0.282	0.614	0.614	0.533	0.553
β**(0.5)**								
Intercept	0.155	0.094	−0.374	−0.375	0.135	0.115	0.213	0.213
Slope	−0.195	−0.242	−0.314	−0.312	0.469	0.471	0.365	0.363
β**(0.7)**								
Intercept	0.237	0.112	−0.307	−0.307	0.206	0.138	0.181	0.180
Slope	−0.309	−0.342	−0.585	−0.583	0.646	0.606	0.656	0.656
				$\mathfrak{n} = 500$				
γ								
Intercept	0.194	0.194	−0.048	−0.047	0.099	0.099	0.055	0.057
Slope	−0.338	−0.338	−0.366	−0.358	0.316	0.316	0.306	0.308
β**(0.5)**								
Intercept	0.136	0.079	−0.369	−0.369	0.062	0.048	0.165	0.165
Slope	−0.177	−0.203	−0.323	−0.323	0.185	0.192	0.202	0.202
β**(0.7)**								
Intercept	0.213	0.112	−0.307	−0.306	0.099	0.062	0.131	0.130
Slope	−0.302	−0.325	−0.575	−0.574	0.290	0.294	0.451	0.451

TABLE 6. Case 5: p = 1, censoring rate = 0.51, cure rate = 0.32; Bias and Mean Squared Error for different estimation methods based on 1000 replications

	LNA	Imp	DA
Intercept	1.069 (0.646)	1.069 (0.646)	0.273 (0.285)
Histology	−0.506 (0.563)	−0.506 (0.563)	−0.471 (0.326)
Age	0.731 (0.312)	0.731 (0.312)	0.591 (0.187)
Sex	−0.686 (0.568)	−0.686 (0.568)	−0.258 (0.376)

TABLE 7. Estimates of the γ parameters for the logistic cure model, bootstrap standard errors in parentheses

and Data Analysis **105,** 144–165.

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clin Cancer Res.* **18,** 3731–3736.

Patilea, V. and van Keilegom, I. (2017). A general approach for cure models in survival analysis. https://arxiv.org/abs/1701.03769v1.

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

Peng, Y. and Dear, K. B. G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56,** 237 – 243.

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

Sy, J. P. and Taylor, J. M. G. (2000). Estimation of a Cox proportional hazards cure model. *Biometrics* **56,**.

Tsodikov, A. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statist. Med.* **21,** 895–920.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* **104,** 1117–1128.

Wang, L., Du, P., and Lian, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics* **68,** 726 – 735.

Wu, Y. and Yin, G. (2013). Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association* **108,** 1517–1531.

Wu, Y. and Yin, G. (2017). Multiple imputation for cure rate quantile regression with censored data. *Biometrics* **73,** 94–103.

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist* **42,** 1 – 17.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Statistical Models of Tumor Latency and their Biostatistical Applications*. World Scientific.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan. *J. Amer. Statist. Assoc.* **87,** 284–292.

Yang, X., Narisetty, N. N., and He, X. (2018). A new approach to censored quantile regression estimation. *Journal of Computational and Graphical Statistics* **18,** 417–425.

Zeileis, A., Koenker, R., and Doebler, P. (2015). *glmx: Generalized Linear Models Extended*. R package version 0.1-1, available from: `https://CRAN.R-project.org/package=glmx`.

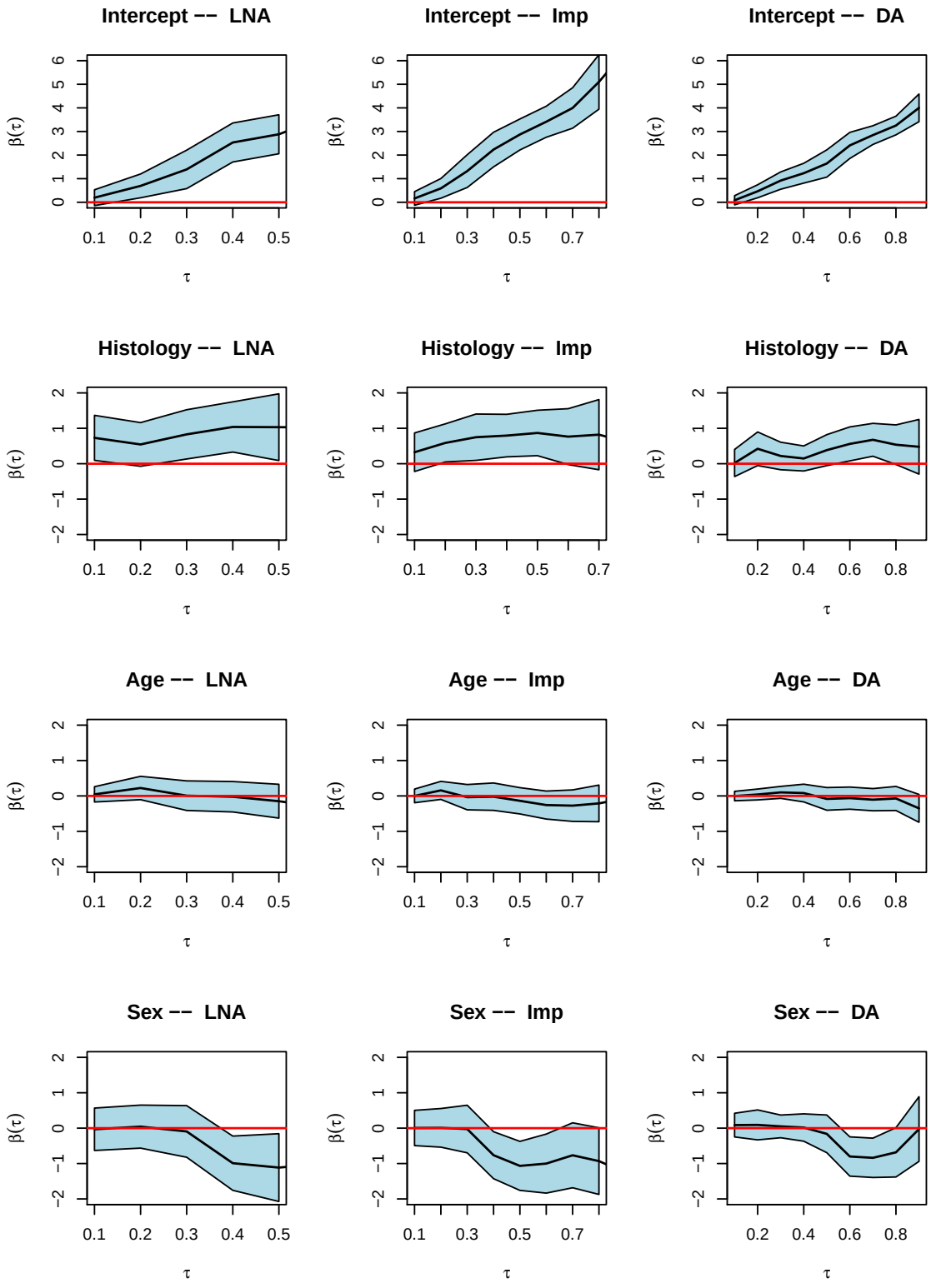

FIGURE 1. Comparison of Three Quantile Regression Estimates for the Lung Cancer Model. The blue pointwise bands in each panel are based on 200 bootstrap replications.