Received: 18 November 2021 | Accepted: 15 July 2022

DOI: 10.1111/biom.13723

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC METHODOLOGY

Quantile regression for nonignorable missing data with its application of analyzing electronic medical records

Aiai Yu[1] | Yujie Zhong[1] | Xingdong Feng[1] | Ying Wei[2]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

[2]Department of Biostatistics, Columbia University, New York, New York, USA

Correspondence
Xingdong Feng, School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Rd, Shanghai 200433, China.
Email: feng.xingdong@mail.sufe.edu.cn

Funding information
National Science Foundation, Grant/Award Number: DMS-1953527; National Natural Science Foundation of China, Grant/Award Numbers: 11690012, 11971292; Foundation for the National Institutes of Health, Grant/Award Numbers: R01-1RF1AG072272, R01-HG008980

Abstract

Over the past decade, there has been growing enthusiasm for using electronic medical records (EMRs) for biomedical research. Quantile regression estimates distributional associations, providing unique insights into the intricacies and heterogeneity of the EMR data. However, the widespread nonignorable missing observations in EMR often obscure the true associations and challenge its potential for robust biomedical discoveries. We propose a novel method to estimate the covariate effects in the presence of nonignorable missing responses under quantile regression. This method imposes no parametric specifications on response distributions, which subtly uses implicit distributions induced by the corresponding quantile regression models. We show that the proposed estimator is consistent and asymptotically normal. We also provide an efficient algorithm to obtain the proposed estimate and a randomly weighted bootstrap approach for statistical inferences. Numerical studies, including an empirical analysis of real-world EMR data, are used to assess the proposed method's finite-sample performance compared to existing literature.

KEYWORDS

estimating equations, inverse probability weight, missing data, Monte Carlo integration, quantile regression

1 | INTRODUCTION

Electronic medical records (EMRs) have been widely used in hospital systems. They contain vast amounts of longitudinal and detailed patient information, including lab tests, medications, disease status, and treatment outcomes. Over the past decade, the EMR systems have become vital research and data resources to understand disease prognosis and real-world treatment responses, and they have led to several successful stories in translational precision medicines (Tatonetti *et al.*, 2012; Doshi-Velez *et al.*, 2014; Li *et al.*, 2015; Miotto *et al.* , 2016).

Analyzing such large, scattered, complex, and heterogeneous patient data for research purposes, however, requires careful considerations. One of its major challenges is handling the large amount of informative missing data in EMRs. Dealing with missing data is a long-standing research topic in statistics. Most methods assume missing at random (Cheng, 1994; Rubin, 1976; van der Laan and McKeague, 1998). However, such an assumption hardly holds in EMRs. In an EMR system, the timing of a patient's hospital visit, as well as which particular variables are recorded after the visit, is dependent on the patient's health conditions, the patient's medical needs, and the clinical judgment of physicians. As a result, the missing mechanism is often related to patients' underlying conditions and is, hence, not missing at random. For example, several clinical studies recommended special ICU management for patients with high glucose levels to reduce their mortality rate (Inzucchi, 2006; Krinsley, 2003, 2004). They may influence physicians' decisions in measuring glucose during ICU admission. The well-known MIMIC-III clinical database includes deidentified EMRs from over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. In the MIMIC-III data, we observed approximately 40% of patients had their glucose levels measured at the time of their ICU admissions, and approximately 30% of observed glucose levels are higher than 180 (ie, hyperglycemia). That is much higher than the hyperglycemia prevalence in general populations. These studies suggest that patients with high glucose levels are more likely to be measured at admission, leading to informative missing data in the glucose levels at admission. EMR analyses without adjusting for such informative missingness are likely to be biased and can hardly be generalized for the population.

Another major challenge of EMRs analysis is identifying and incorporating data heterogeneity and association complexity. Most human disease progressions are complex, and their associations with exposures and biomarkers are highly heterogeneous in a population. The traditional mean-based regressions seldom fully capture such heterogeneity and complexity in EMR data. There has been a growth of interest in investigating the distributional effect of exposures on an outcome of interest (Bind *et al.*, 2015, 2016; Gong *et al.*, 2020; Wu *et al.*, 2021). Quantile regression, proposed by Koenker and Bassett (1978), is an approach to learn distributional associations by estimating the conditional quantiles of the response variable Y given covariates $\mathbf{X}$. Since people with health risks often present either high or low values of biomarkers, modeling upper or lower quantiles naturally stratifies the population by their health risks and provides direct insights into the associations among subpopulations at risk. In the analysis presented in the Application section, the proposed quantile analysis discovered strong social–demographic heterogeneity at the tails of glucose distributions, which would have been underestimated in traditional mean-based analyses.

In addition, quantile regression enjoys robustness against response outliers due to its bounded subgradient of the loss function, another desired property for analyzing error-prone EMR data. Even when conditional means are of interest, several studies reported that aggregating quantile regression across quantile levels could be more efficient than direct mean regressions in estimating conditional means and other summary statistics when data are nonnormally distributed, especially when data contain heavy tails (ie, outliers) (Zou and Yuan, 2008; Zhao and Xiao, 2014). In the same MIMIC-III data, the observed glucose values at admission range from 30 mg/dL to 923 mg/dL, indicating the existence of outliers in these glucose measurements.

To address those challenges in EMR data, we propose a new estimation and inference framework to enable unbiased quantile regression with nonignorable missing responses for EMR analyses. Classical approaches that handle missing data assume missing at random and use multiple imputation (Aerts *et al.*, 2002; Lipsitz *et al.*, 1998; Rubin, 1978), or the inverse probability weighting (Horvitz and Thompson, 1952; Robins *et al.*, 1994, 1995), to adjust for the missing data. When missing data are nonignorable, estimations become a lot more challenging. Statistical tools for nonignorable missing data have been very limited until a recent methodological advancement. Kim and Yu (2011) proposed an inverse propensity weighting approach with a known exponential tilting propensity function. Shao and Wang (2016) further proposed an instrumental variable method to estimate such propensity. Tang *et al.* (2003) proposed pseudo-likelihood estimators for multivariate responses with nonignorable missing data, and Zhao and Ma (2018) further generalized and established the asymptotic properties of the pseudo-likelihood estimators. Other approaches include the propensity-score-adjusted imputation method in Zhao *et al.* (2017) and the calibration method in Han (2018). Those methods are based on mean regression models and are not directly applicable to quantile regression analysis. There has been minimal literature for quantile regression with nonignorable missing. Yuan and Yin (2010) and Ghasemzadeh *et al.* (2018) considered Bayesian approaches for longitudinal outcomes. Zhao *et al.* (2017) studied several inverse probability weighting estimators for quantile regression coefficients, and Zhang and Wang (2020) developed a smoothed weighted empirical likelihood method based on inverse probability weighting and kernel smoothing approaches. All of these approaches are either considered parametric models that are not compatible with the quantile regression model, or they only use completely observed data. The former leads to biased estimates, while the latter does not optimize the estimation efficiency.

This motivates us to develop a novel inverse probability weighting method for the quantile regression estimation with nonignorable missing responses. Our approach is fully compatible with both quantile regression and the missing mechanism model to mitigate potential bias and use all of the observed data to improve estimation efficiency. The rest of this paper is organized as follows. Section 2 introduces the estimation method, and proposes an iterative algorithm to obtain estimates. The asymptotic properties of the proposed estimator are given in Section 3, along with a randomly weighted bootstrap method. In Section 4, a simulation study is conducted to examine the finite-sample performance of the proposed method. An application to the motivating study on EMRs is provided in Section 5. Finally, Section 6 concludes the paper. All proofs of theorems are deferred to the web-based supplementary materials in the Supporting Information section.

2 | METHOD

2.1 | Model setting

We consider the following quantile regression model:

$$Q_Y(\tau \mid \mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}_{0,\tau}, \quad \text{for } \tau \in (0,1), \tag{1}$$

where Y is an outcome of interest, $\mathbf{X}$ is the p-dimensional vector of the covariate including the intercept, τ is the quantile level, and $Q_Y(\tau \mid \mathbf{X})$ stands for the τth conditional quantile of Y given $\mathbf{X}$. We assume that model (1) holds for all $\tau \in (0,1)$, and we denote $F(y \mid X)$ as the conditional distribution derived from model (1).

Suppose $(y_i, \mathbf{x}_i^\top)$ is an independent and identically distributed random sample satisfying $F(y \mid \mathbf{X})$. If all data are completely observed, the parameter $\boldsymbol{\beta}_{0,\tau}$ can be consistently estimated by $\widehat{\boldsymbol{\beta}}_\tau = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$, where $\rho_\tau(u) = \{\tau - I(u<0)\}u$ is quantile regression loss function (Koenker, 2005). Or, equivalently, one can estimate $\boldsymbol{\beta}_{0,\tau}$ by solving the estimating equations $\sum_{i=1}^n \mathbf{x}_i \Psi_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) = 0$, where $\Psi_\tau(u) = I(u<0) - \tau$ is the piecewise first derivative of ρ_τ with $I(\cdot)$ standing for an indicator function. When some responses y_i are missing *not* at random, for instance, the probability of missingness may depend upon the value of y_i itself (Rubin, 1976), then the estimate $\widehat{\boldsymbol{\beta}}_\tau$ from the observed data is biased.

Missing Mechanism Model.
Let δ_i be the binary indicator of y_i being observed. We assume the following model for δ_i,

$$pr(\delta = 1 \mid \mathbf{x}, y) = pr(\delta = 1 \mid \mathbf{x}_1, y) = \pi(\mathbf{x}_1, y; \boldsymbol{\theta}_0), \tag{2}$$

where $\mathbf{x}_1$ is a d-dimensional subset of the covariate $\mathbf{x}$ and $\pi(\cdot)$ is a parametric function with $(d+1)$-dimensional parameter $\boldsymbol{\theta}_0$. Although we consider the parametric missing mechanism, it can be generalized with the semiparametric methods as suggested by Kim and Yu (2011) and Shao and Wang (2016). We denote the $(p-d)$-dimensional complementary covariate as $\mathbf{x}_2$. Model (2) implies that the missing mechanism is nonignorable in the sense that the missingness of response is not independent of the response variable y, even after the effect of the covariate $\mathbf{x}$ is taken into account. Besides, models (1) and (2) imply that the covariate $\mathbf{x}_2$ is associated with the response y but independent of the missingness process that is conditional on the covariate $\mathbf{x}_1$ and response y, that is, $\mathbf{x}_2\ nupmodelsy \mid \mathbf{x}_1$ and $\mathbf{x}_2\ upmodels\delta \mid (\mathbf{x}_1, y)$. It follows from Wang *et al.* (2014) and Miao and Tchetgen (2018) that the parameters in models (1) and (2) are identifiable. In the rest of paper, we call $\mathbf{x}_2$ the instrumental variable following the conventions in literature.

2.2 | Estimation

Following the definition of $\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta}_0)$, one can obtain a consistent estimator of $\boldsymbol{\beta}_{0,\tau}$ by solving the following estimating equations for $\boldsymbol{\beta}$:

$$\mathbf{M}_{n,\tau}(\boldsymbol{\theta}_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta}_0)} \Psi_\tau\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)\mathbf{x}_i = 0.$$

Since $\boldsymbol{\theta}_0$ are unknown parameters, we construct the following unbiased augmented likelihood for estimating $\boldsymbol{\theta}_0$:

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \prod_{i=1}^n L_{ni}(\boldsymbol{\theta}) \\ &= \prod_{i=1}^n \{\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta})\}^{\delta_i} \left[\mathrm{E}_y\{1 - \pi(\mathbf{x}_{1i}, y; \boldsymbol{\theta}) \mid \mathbf{x}_i\}\right]^{1-\delta_i}, \end{aligned}$$

where $\mathrm{E}_y\{1 - \pi(\mathbf{x}_{1i}, y; \boldsymbol{\theta}) \mid \mathbf{x}_i\} = \int_y \{1 - \pi(\mathbf{x}_{1i}, y; \boldsymbol{\theta})\} f(y \mid \mathbf{x}_i) dy$ is the marginal probability of $\delta_i = 0$ given the observed $\mathbf{x}_i$. Apparently, the unbiased estimation of $\boldsymbol{\theta}_0$ relies on a correctly specified density $f(y \mid \mathbf{x})$ to evaluate the conditional expectations in $L_n(\boldsymbol{\theta})$. Such density satisfies the quantile regression model (1) to ensure model compatibility. Parametric models, such as those commonly used as those of Greenlees *et al.* (1982) and Riddles *et al.* (2016), often fail and lead to biased estimates of $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_{0,\tau}$. We will introduce an alternative joint estimating equation approach that consistently estimates $\boldsymbol{\beta}_{0,\tau}$ and $\boldsymbol{\theta}_0$ based on models (1) and (2) without further assumptions on $f(y \mid \mathbf{x})$.

Let $\boldsymbol{\beta}_0(\tau)$ be the true quantile regression process of model (1) with $\boldsymbol{\beta}_{0,\tau} = \boldsymbol{\beta}_0(\tau)$. Model (1) implies that the conditional quantile function of Y given $\mathbf{x}$ is $\mathbf{x}^\top \boldsymbol{\beta}_0(\tau)$, and its corresponding conditional probability density $f(y \mid \mathbf{x})$ can be written as

$$f(y \mid \mathbf{x}) = \left[\frac{\partial\{\mathbf{x}^\top \boldsymbol{\beta}_0(\tau)\}}{\partial \tau}\right]^{-1}\Bigg|_{\tau = \tau_y\{\boldsymbol{\beta}_0(\tau)\}},$$

where $\tau_y\{\boldsymbol{\beta}_0(\tau)\} = \inf\{\tau \in (0,1) : \mathbf{x}^\top \boldsymbol{\beta}_0(\tau) > y\}$ is the quantile level of y with respect to the quantile function $\mathbf{x}^\top \boldsymbol{\beta}_0(\tau)$. It is clear that the conditional density $f(y \mid \mathbf{x})$ depends on the unknown quantile regression process $\boldsymbol{\beta}_0(\tau)$. We hence expand the likelihood L_n over the joint parameter space of $(\theta, \boldsymbol{\beta}(\tau))$ by

$$\begin{aligned}&L_n(\theta, \boldsymbol{\beta}(\tau))\\ &= \prod_{i=1}^{n} \{\pi(\mathbf{x}_{1i}, y_i; \theta)\}^{\delta_i} \left[\int_y \{1-\pi(\mathbf{x}_{1i}, y; \theta)\} f(y \mid \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)) dy\right]^{1-\delta_i},\end{aligned}$$

where $f(y \mid \mathbf{x}_i^\top \boldsymbol{\beta}(\tau))$ is the conditional density induced from the quantile process $\mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$. When $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$, $f(y \mid \mathbf{x}_i^\top \boldsymbol{\beta}(\tau))$ is the true conditional density.

With these notations, we construct the following joint estimating equations for $(\theta_0, \boldsymbol{\beta}_0(\tau))$:

$$\mathbf{M}_n(\theta, \boldsymbol{\beta}(\tau)) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi(\mathbf{x}_{1i}, y_i; \theta)} \Psi_\tau\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}(\tau)\right)\mathbf{x}_i = 0, \quad (3)$$

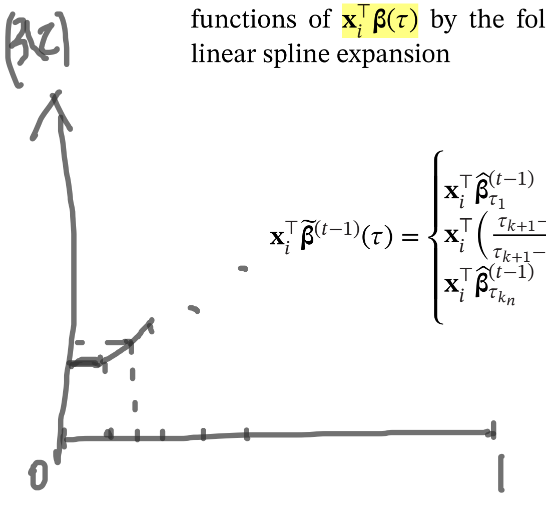
$$\mathbf{S}_n(\theta, \boldsymbol{\beta}(\tau)) = \frac{1}{n}\frac{\partial \log L_n(\theta, \boldsymbol{\beta}(\tau))}{\partial \theta} = 0. \qquad (4)$$

2.3 | Computing algorithm

In reality, one cannot estimate $\boldsymbol{\beta}_{0,\tau}$ for all $\tau \in (0,1)$. Instead, we solve the estimation equations (5) over an evenly spaced fine grid of quantile levels $\Omega = \{\epsilon_n = \tau_1 < \dots < \tau_{k_n} = 1 - \epsilon_n\}$, where ϵ_n is a small positive number that may depend on the sample size n. Denoting $\boldsymbol{\phi} = (\boldsymbol{\beta}_{\tau_1}^\top, \boldsymbol{\beta}_{\tau_2}^\top, \dots, \boldsymbol{\beta}_{\tau_{k_n}}^\top)^\top$, and letting $\boldsymbol{\phi}_0$ be the corresponding true quantile coefficients, we propose an iterative algorithm to obtain unbiased estimates of $\boldsymbol{\xi}_0 = (\theta_0^\top, \boldsymbol{\phi}_0^\top)^\top$. We start with initializing parameters $\boldsymbol{\phi}$ by quantile regression with complete observations, and we denote the resulting estimates as $\widehat{\boldsymbol{\phi}}^{(0)} = (\widehat{\boldsymbol{\beta}}_{\tau_1}^{(0)\top}, \dots, \widehat{\boldsymbol{\beta}}_{\tau_{k_n}}^{(0)\top})^\top$. Letting t be indicator of iteration steps, we outline the algorithm as follows and refer to Web Appendix A of the Supporting Information for the details.

Step 1. Suppose $\widehat{\boldsymbol{\phi}}^{(t-1)} = (\widehat{\boldsymbol{\beta}}_{\tau_1}^{(t-1)\top}, \dots, \widehat{\boldsymbol{\beta}}_{\tau_{k_n}}^{(t-1)\top})^\top$ is the estimated quantile coefficients at the $(t-1)$th iteration, then approximate the conditional quantile functions of $\mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$ by the following piecewise linear spline expansion

$$\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(t-1)}(\tau) = \begin{cases} \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{\tau_1}^{(t-1)} & \tau < \tau_1 \\ \mathbf{x}_i^\top \left(\frac{\tau_{k+1}-\tau}{\tau_{k+1}-\tau_k}\widehat{\boldsymbol{\beta}}_{\tau_k}^{(t-1)} + \frac{\tau-\tau_k}{\tau_{k+1}-\tau_k}\widehat{\boldsymbol{\beta}}_{\tau_{k+1}}^{(t-1)}\right) & \tau \in [\tau_k, \tau_{k+1}) \text{ for } k = 1, 2, \dots, k_n - 1. \\ \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{\tau_{k_n}}^{(t-1)} & \tau \geq \tau_{k_n} \end{cases}$$



Step 2. Use Monte Carlo average to approximate the conditional expectations in $\mathbf{S}_n(\theta, \boldsymbol{\beta}(\tau))$, for example, $\mathrm{E}_y\{1 - \pi(\mathbf{x}_{1i}, y; \theta) \mid \mathbf{x}_i\} \approx \sum_{l=1}^{m}\{1 - \pi(\mathbf{x}_{1i}, \widetilde{y}_i^l; \theta)\}/m$, where $\widetilde{y}_i^l, l = 1, \dots, m$ are m random samples drawn from the quantile function $\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(t-1)}(\tau)$.

Step 3. Update $\widehat{\theta}^{(t)}$ by solving the equation $\mathbf{S}_n(\theta, \widetilde{\boldsymbol{\beta}}^{(t-1)}(\tau)) = 0$.

Step 4. Update $\widehat{\boldsymbol{\phi}}^{(t)}$ by solving the equations $(\mathbf{M}_{n,\tau_1}^\top(\widehat{\theta}^{(t)}, \boldsymbol{\beta}_{\tau_1}), \dots, \mathbf{M}_{n,\tau_{k_n}}^\top(\widehat{\theta}^{(t)}, \boldsymbol{\beta}_{\tau_{k_n}}))^\top = 0$.

Step 5. Repeat Steps 1 to 4 until some stopping criteria are satisfied. We denote the resulting estimates $\widehat{\boldsymbol{\xi}}_n = (\widehat{\theta}_n^\top, \widehat{\boldsymbol{\phi}}_n^\top)^\top$.

Remark 1. In Step 1, as pointed out by Wei and Carroll (2009), the approximation process $\mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}^{(t-1)}(\tau)$ uniformly converges to the true conditional quantile function as the number of knots k_n increases, under certain mild conditions.

Remark 2. Due to the structure of Monte Carlo integration, Step 3 is a weighted logistic regression, where the response is δ_i, the observations $(\mathbf{x}_{1i}, y_i)$ are weighted with the constant one if $\delta_i = 1$, and the simulated pairs $(\mathbf{x}_{1i}, \widetilde{y}_i^l)$ are weighted with the ratios $\{1 - \pi(\mathbf{x}_{1i}, \widetilde{y}_i^l; \widehat{\theta}^{(t-1)})\}/\sum_{l=1}^{m}\{1 - \pi(\mathbf{x}_{1i}, \widetilde{y}_i^l; \widehat{\theta}^{(t-1)})\}$ if $\delta_i = 0$.

3 | ASYMPTOTIC PROPERTIES

3.1 | Limiting behavior

In this section, we establish the asymptotic results for the estimate $\widehat{\boldsymbol{\xi}}_n$, and we first introduce some assumptions.

Assumption 1. The covariate $\mathbf{x}$ is subexponential.

Assumption 2. The true coefficient quantile function $\boldsymbol{\beta}_0(\cdot)$ is smooth and differentiable for any $\tau \in (0,1)$. Let $h_\mathbf{x}(\tau) = 1/\mathbf{x}^\top \boldsymbol{\beta}_0'(\tau)$, which is the density of y given $\mathbf{x}$ at the τth quantile. For any $\mathbf{x} \in \mathcal{X}$,

(i) $0 < h_\mathbf{x}(\tau) < \infty$ and $\lim_{\tau \to 0} h_\mathbf{x}(\tau) = \lim_{\tau \to 1} h_\mathbf{x}(\tau) = 0$;

(ii) there exist constants M and $v_1, v_2 > -1$ such that its first derivative is bounded by $\sup_{\mathbf{x}} h'_{\mathbf{x}}(\tau) < M\tau^{v_1}(1-\tau)^{v_2}$.

Assumption 3. $\pi(\mathbf{x}_1, y; \theta)$ is twice differentiable with respect to θ; $L_\pi \le \pi(\mathbf{x}_1, y; \theta) \le U_\pi$ holds almost surely for some positive constants $0 < L_\pi < U_\pi < 1$; $\partial_\theta \pi(\mathbf{x}_1, y; \theta) := \partial\pi(\mathbf{x}_1, y; \theta)/\partial\theta$ is uniformly bounded.

Remark 3. The subexponential family includes many commonly used distributions, such as Gaussian, Gamma, and Poisson distributions. Assumption 2 is also considered by Wei and Carroll (2009), which ensures that the true coefficient function $\boldsymbol{\beta}_0(\cdot)$ is smooth enough to be well approximated by natural linear splines. Assumption 2(i) implicitly assumes that the conditional density $f(y \mid \mathbf{x})$ is continuous and bounded away from both zero and infinity, and that it diminishes to zero as the quantile level τ goes to 0 or 1. Assumption 2(ii) holds for a wide range of distribution families, including those in the exponential family. Assumption 3 imposes the restriction on the missing probabilities, which is commonly considered in the literature (Kim and Riddles, 2012; Qin *et al.*, 2008; Zhao *et al.*, 2017).

Denote $\mathbf{H}^0_{n,\tau}(\theta, \boldsymbol{\beta}) = (\mathbf{S}_n^\top(\theta), \mathbf{M}_{n,\tau}^\top(\theta, \boldsymbol{\beta}))^\top$, $\mathbf{H}_n^0(\boldsymbol{\xi}) = (\mathbf{S}_n^\top(\theta), \mathbf{M}_n^\top(\boldsymbol{\xi}))^\top$, and $\mathbf{H}_n(\boldsymbol{\xi}) = (\mathbf{S}_n^\top(\theta, \widetilde{\boldsymbol{\beta}}(\tau)), \mathbf{M}_n^\top(\boldsymbol{\xi}))^\top$, where $\mathbf{M}_n(\boldsymbol{\xi}) = (\mathbf{M}^\top_{n,\tau_1}(\theta, \boldsymbol{\beta}_{\tau_1}), \dots, \mathbf{M}^\top_{n,\tau_{k_n}}(\theta, \boldsymbol{\beta}_{\tau_{k_n}}))^\top$.

Assumption 4. The true coefficient $(\theta_0^\top, \boldsymbol{\beta}_{0,\tau}^\top)^\top$ is the unique solution to the equation $\mathrm{E}\{\mathbf{H}^0_{n,\tau}(\theta, \boldsymbol{\beta})\} = 0$ for $\tau \in (0, 1)$. There also exists a unique solution $\boldsymbol{\xi}^\star = (\theta^{\star\top}, \boldsymbol{\phi}^{\star\top})^\top$ to the equation $\mathrm{E}\{\mathbf{H}_n(\boldsymbol{\xi})\} = 0$.

Assumption 5. There exist some compact sets $\Theta \in \mathbb{R}^{d+1}, \Theta_{\boldsymbol{\phi}} \in \mathbb{R}^{k_n p}$ such that $\theta_0 \in \Theta$, $\boldsymbol{\phi}_0 \in \Theta_{\boldsymbol{\phi}}$, respectively.

Remark 4. Assumption 4 is the identifiability condition commonly used in the literature (Wei and Carroll, 2009; Wei and Yang, 2014). Assumption 5 is also commonly used for the mathematical development in establishing the consistency of estimators (Chen *et al.*, 2015; Wang and Feng, 2012).

Theorem 1. *Under Assumptions 1–5, if $n \to \infty$, $k_n \to \infty$, $k_n/n \to 0$, and $m \to \infty$, we have*

$$\left\|\widehat{\theta}_n - \theta_0\right\| = o_p(1), \text{ and } \sup_{\tau \in [1/(k_n+1), k_n/(k_n+1)]} \left\|\widehat{\boldsymbol{\beta}}_n(\tau) - \boldsymbol{\beta}_0(\tau)\right\| = o_p(1).$$

The detail of the proof of Theorem 1 is given in Web Appendix B.

We further denote $\partial^2_\theta \pi(\mathbf{x}_1, y; \theta) := \partial^2\pi(\mathbf{x}_1, y; \theta)/\partial\theta\partial\theta^\top$, $\Theta_{\boldsymbol{\xi}} = \Theta \times \Theta_{\boldsymbol{\phi}}$, and let $\lambda_{\max}(A), \lambda_{\min}(A)$ be the maximum and minimum absolute values of the eigenvalues of the matrix A, respectively.

X		
1	0	0
2	1	0
3	0	1

Assumption 6. For all $\theta \in \Theta$, $\sum_{j=1}^{d+1} \|\{\partial^2_\theta \pi(\mathbf{x}_1, y; \theta)\}_{(.,j)}\|$ is bounded by an square integrable function, where $A_{(.,j)}$ denotes the j-th column of the matrix A.

Assumption 7. $v = \min\{v_1, v_2\} > -1/2$, where v_1, v_2 are the constants given in Assumption 2(ii).

Assumption 8. There exists a sequence of $q_n \times q_n$ matrices D_n with $\liminf_{n\to\infty} \lambda_{\min}(D_n^\top D_n) > 0$, where $q_n = k_n p + d + 1$, such that, for some $\epsilon > 0$ and $\gamma > 0$,

$$\sup_{\boldsymbol{\xi} \in \Theta_{\boldsymbol{\xi}} : \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| \le \epsilon} \frac{\|\mathrm{E}\{\mathbf{H}_n^0(\boldsymbol{\xi})\} - \mathrm{E}\{\mathbf{H}_n^0(\boldsymbol{\xi}_0)\} - D_n(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\|}{\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|^{1+\gamma}} < \infty.$$

Remark 5. Assumption 6 imposes the smoothness condition on the missing distribution density $\pi(\mathbf{x}_1, y; \theta)$. Although a parametric missing model is considered, we expect the resulting quantile estimates to remain consistent with more generalized missing mechanism models. More careful derivations are required to determine the proper convergence rate, limiting distributions, and resampling adjustment, similarly to those given in Theorems 2–3. Assumption 7 further limits the distribution space of responses compared with Assumption 2(ii), which still contains many commonly used distributions. Assumption 8 is used to obtain the asymptotic covariance of the proposed estimator (He and Shao, 2000).

Theorem 2. *Under Assumptions 1–8, if $n, m \to \infty$, $k_n^2 \log^2(n)/n \to 0$, $k_n^{3+2v}/n \to \infty$, and $k_n m/n \to \infty$, then for any $\boldsymbol{\alpha} \in \mathbb{R}^{q_n}$, $\|\boldsymbol{\alpha}\| = 1$, we obtain that*

$$\frac{\sqrt{n}\boldsymbol{\alpha}^\top\left(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\right)}{\sigma(\boldsymbol{\alpha})} \xrightarrow{L} \mathcal{N}(0, 1), \quad n, m \to \infty,$$

where $\sigma^2(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top D_n^{-1} V_n (D_n^{-1})^\top \boldsymbol{\alpha}$, $V_n = \mathrm{var}\{n^{1/2}\mathbf{H}_n^0(\boldsymbol{\xi}_0)\}$, *and D_n is given in Assumption 8.*

Remark 6. When there are no missing data, that is $\pi(\mathbf{x}_i, y_i; \theta_0) \equiv 1$, we have $\mathbf{S}_n(\theta, \widetilde{\boldsymbol{\beta}}(\tau)) = \mathbf{S}_n(\theta) = 0$. If we consider a single quantile level ($k_n = 1$), then $\boldsymbol{\xi} = \boldsymbol{\beta}_\tau$ and $\mathbf{H}_n(\boldsymbol{\xi}) = \mathbf{H}_n^0(\boldsymbol{\xi}) = \sum_{i=1}^n \Psi_\tau(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}_\tau)\mathbf{x}_i/n$. In this case, the result of Theorem 2 coincides with Theorem 4.1 of Koenker (2005). Furthermore, if data are missing at random, that is $\pi(\mathbf{x}_i, y_i; \theta_0) = \pi(\mathbf{x}_i; \theta_0)$, we then have $\mathbf{S}_n(\theta, \widetilde{\boldsymbol{\beta}}(\tau)) = \mathbf{S}_n(\theta)$, $\boldsymbol{\xi} = (\theta^\top, \boldsymbol{\beta}_\tau^\top)^\top$, and $\mathbf{H}_n(\boldsymbol{\xi}) = \mathbf{H}_n^0(\boldsymbol{\xi}) = (\mathbf{S}_n^\top(\theta), \mathbf{M}_{n,\tau}^\top(\theta, \boldsymbol{\beta}_\tau))^\top$ when we still focus on a single quantile level. Then, the result of Theorem 2 coincides with that of Sherwood *et al.* (2013).

$Z_1 = \begin{cases} 1 & 白 \\ 0 & 非白 \end{cases}$ $Z_2 = \begin{cases} 1 & 黑 \\ 0 & 非黑 \end{cases}$

$Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon .$

The detail of the proof of Theorem 2 is deferred to Web Appendix C. Theorem 2 establishes the asymptotic distribution of the proposed estimator. However, it is challenging to directly estimate the asymptotic variance due to its complicated form. Thus, we consider a resampling method in the following subsection.

$x_2 \in \{1, 2, 3\}$

白 黑 西

3.2 | Resampling method

Adopting the idea of the method of Rao and Zhao (1992), we construct a resampling approach for statistical inferences with the proposed estimates. We use a sequence of random weights $\{w_1, w_2, \dots, w_n\}$ which satisfy the following assumption.

Assumption 9. The random weights $w_1, w_2, \dots, w_n$ are independent and identically generated with $pr(w_1 > 0) = 1$, $\mathrm{E}(w_1) = \mathrm{var}(w_1) = 1$.

Let $\widehat{\mathbf{H}}_{w,n}(\boldsymbol{\xi}) = (\widehat{\mathbf{S}}_{w,n}^\top(\boldsymbol{\xi}), \mathbf{M}_{w,n}^\top(\boldsymbol{\xi}))^\top$, where

$$\widehat{\mathbf{S}}_{w,n}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n} w_i \left[\delta_i \frac{\partial_{\boldsymbol{\theta}}\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta})}{\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta})} - (1-\delta_i)\frac{\sum_{l=1}^{m} \partial_{\boldsymbol{\theta}}\pi(\mathbf{x}_{1i}, \tilde{y}_i^l; \boldsymbol{\theta})}{\sum_{l=1}^{m}\{1 - \pi(\mathbf{x}_{1i}, \tilde{y}_i^l; \boldsymbol{\theta})\}}\right],$$

$$\mathbf{M}_{w,n}(\boldsymbol{\theta}, \boldsymbol{\beta}_\tau) = \frac{1}{n}\sum_{i=1}^{n} w_i \frac{\delta_i}{\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta})}\Psi_\tau\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\tau\right)\mathbf{x}_i,$$

and $\mathbf{M}_{w,n}(\boldsymbol{\xi}) = (\mathbf{M}_{w,n}^\top(\boldsymbol{\theta}, \boldsymbol{\beta}_{\tau_1}), \dots, \mathbf{M}_{w,n}^\top(\boldsymbol{\theta}, \boldsymbol{\beta}_{\tau_{k_n}}))^\top$. Then, the bootstrap estimate of the parameter $\boldsymbol{\xi}_0$, denoted as $\widehat{\boldsymbol{\xi}}_n^*$, can be obtained by solving $\widehat{\mathbf{H}}_{w,n}(\boldsymbol{\xi}) = 0$.

Let pr^* denote the probability induced by the bootstrap method given the data. We then have the following result, which validates the proposed resampling inference tool.

Theorem 3. *Under Assumptions 1–9, if the conditions of Theorem 2 hold, then for any $\alpha \in \mathbb{R}^{q_n}$, $\|\alpha\| = 1$, we obtain that*

$$\sup_{z \in \mathbb{R}} \left| pr^*\left\{ n^{1/2}\alpha^\top\left(\widehat{\boldsymbol{\xi}}_n^* - \widehat{\boldsymbol{\xi}}_n\right)\Big/\sigma(\alpha) < z \right\} - pr\left\{ n^{1/2}\alpha^\top\left(\widehat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\right)\Big/\sigma(\alpha) < z \right\}\right|$$

converges in probability to zero, where $\sigma(\alpha)$ is provided in Theorem 2.

Theorem 3 provides the theoretical justification for the randomly weighted bootstrap method, and the detail of the proof is provided in Web Appendix D.

4 | SIMULATION STUDY

In this section, we conduct a simulation study to assess the finite-sample performance of the proposed method in comparison to the existing approaches in the literature. In what follows, we present the designed simulation settings, the methods under comparison, and the metrics for assessing estimation performances, followed by an in-depth discussions of the results and comparisons.

Simulation Models.
We first consider the following location-scale model for the outcome

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + (\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i})\varepsilon_i, \quad i = 1, \dots, n, \tag{5}$$

where $(\beta_0, \beta_1, \beta_2, \beta_3) = (1, -2, 2, 0.5)$, and $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.5, 0.5, 0.2, 0)$. In the location-scale model, both intercept and slope change with quantile levels. We consider two sample sizes $n = 500, 1000$, and generate the covariates x_{1i} and x_{2i} from the Bernoulli distribution with the success probability 0.5 and the normal distribution $\mathcal{N}(2, 0.5^2)$, respectively. Furthermore, to introduce correlations among covariates, we generate $x_{3i} = 0.3x_{2i} + u_i$ with the random variable u_i generated from the uniform distribution $U(0, 1)$. We choose the coefficient 0.3 to make the correlation coefficient between x_{2i} and x_{3i} close to 0.5. Under model (13), the corresponding τth conditional quantile of y_i is then given by $Q_{y_i}(\tau \mid x_i) = \beta_0(\tau) + \beta_1(\tau)x_{1i} + \beta_2(\tau)x_{2i} + \beta_3(\tau)x_{3i}$, where $\beta_0(\tau) = \beta_0 + 0.5Q_\varepsilon(\tau)$, $\beta_1(\tau) = \beta_1 + 0.5Q_\varepsilon(\tau)$, $\beta_2(\tau) = \beta_2 + 0.2Q_\varepsilon(\tau)$ and $\beta_3(\tau) \equiv 0.5$ with $Q_\varepsilon(\tau)$ being the τth quantile of the error ε. In this study, we consider the following three distributions with different skewness for the random error ε.

Setting 1: $\varepsilon \sim \mathcal{N}(0, 0.5^2)$;
Setting 2: $\log(\varepsilon + 1) \sim \mathcal{N}(-0.25, 0.5)$;
Setting 3: $\varepsilon \sim \{\chi^2(3) - 3\}/\sqrt{6}$.

We generate the missing indicator δ_i (for the response y_i) from a conditional Bernoulli distribution with the missing probability specified by $\pi(\mathbf{x}_{1i}, y_i; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 y_i)/\{1 + \exp(\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 y_i)\}$, where $\mathbf{x}_{1i} = (1, x_{1i}, x_{2i})^\top$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)^\top$. We consider two parameter settings, $\boldsymbol{\theta} = (-2, 0.5, 0.5, 0.5)^\top$ and $(-3, 0.5, 0.5, 0.5)^\top$, which result in 20% and 40% missing responses, respectively.

Methods under comparison.
For each random sample generated from a simulation setting (determined by error distribution, missing parameter, and sample size), we apply the following six methods to

estimate the coefficient $\boldsymbol{\beta}_\tau$ at the quantile levels $\tau = 0.1$, 0.25, 0.50, 0.75, and 0.9.

(a) The oracle method assuming all data are observed (fullQr);
(b) The naive method based on complete cases only (NaiveQr);
(c) The parametric method of Zhao *et al.* (2017) with the assumption that the response y given the covariate $\mathbf{x}$ and the indicator $\delta = 1$ is normally distributed (elmIpwQr);
(d) The semiparametric method of Zhao *et al.* (2017) without any parametric specifications on the unconditional joint distribution of $(\mathbf{x}, y)$ (elsIpwQr);
(e) The smoothed weighted empirical likelihood method of Zhang and Wang (2020) (swelQr);
(f) The proposed method with $k_n = 20$ evenly spaced quantile grids and $m = 100$ random samples drawn from the quantile function $\mathbf{x}^\top \widetilde{\boldsymbol{\beta}}(\tau)$ for each missing response y_i (proIpwQr).

Metrics for estimation performance.
We assess the finite-sample estimation performance in terms of the mean biases and the square root of mean square errors (RMS). To evaluate the proposed resampling confidence intervals (CIs), we report the empirical coverage probabilities of CIs with the nominal level 95%, denoted as 95ECP, together with their lengths. To construct CIs based on normal approximations, we use 100 resamples for each method with the random weights independently generated from the exponential distribution exp(1) except for the method swelQr where the approach suggested by Zhang and Wang (2020) is used. To conserve space, we only report the results of Setting 3 with the sample size $n = 500$ and 40% missing rate in the paper, and the rest is summarized in Web Appendix E.1. In addition, we conduct numerical studies to explore the effect of the turning parameters k_n and m on the performance of the proposed proIpwQr, and we then compare the average computing time of different methods under the three considered settings. Their numerical results can be found in Web Appendices E.2 and E.3, respectively.

Results and Comparisons.
We report the resulting mean biases, RMS, 95ECP, and CI-length from 1000 Monte Carlo replicates from the six methods under Setting 3 in Table 1. Those under Settings 1 and 2 are included in the Supporting Information. As shown in each of these tables, the naive method leads to substantive biases, especially when the distribution of the error ε is nonnormal or when the missing rate is high. Due to the underlying model misspecification in elmIpwQr, its estimates are more biased than those from elsIpwQr, swelQr, and proIpwQr in all three settings. When the data are normally distributed (Setting 1), the performances of elsIpwQr, swelQr, and proIpwQr are comparable in point estimation and CI construction. However, when the data are skewed (Setting 2) or contain heavier tails (Setting 3), the proposed method proIpwQr produces a much smaller bias and more accurate CI coverage than other approaches. Using Setting 3 as an example, we found that propIpwQr has the slightest bias and mean square errors in all the slope estimates. When it comes to CI construction, the empirical coverage probabilities of the resampling CI from propIpwQr are closest to their nominal levels with short lengths. In comparison, elsIpwQr tends to undercover the true parameters, while swelQr overestimates the estimation errors, leading to a longer length of CI. Hence, we conclude that the proposed method performs well in correcting the biases that result from informative missing data, and it can deliver accurate quantile estimations and inferences under various settings.

5 | EMPIRICAL ANALYSIS OF EMRs

5.1 | Analysis of MIMIC-III clinical data

This section applies the proposed method to a part of MIMIC-III clinical data, a large and widely used real-world clinical database. The data set includes deidentified EMRs associated with more than 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson *et al.*, 2016a, 2016b).

The database contains detailed patient-level information, including demographics, laboratory test results, procedures, medications, mortality, and other clinical information. It has supported a wide range of clinical and epidemiological research since being freely available to researchers worldwide.

Patients from underserved communities often have difficulties managing chronic health conditions due to limited access to health resources. This leads to higher morbidity and mortality rates. We use a part of MIMIC-III data to investigate how sociodemographic and socioeconomic determinants impact glucose control in a real-world setting. In particular, we use linear quantile regression models to analyze how the distributions of glucose levels at admission vary by several sociodemographic and socioeconomic determinants, including age, gender, race, insurance, and marital status. Both hyperglycemia (blood sugar level too high) and hypoglycemia (blood sugar level too low) are common ICU complications and linked to increased morbidity and mortality (Baker *et al.*, 2020). By modeling both upper and lower tails of glucose levels,

TABLE 1 (Setting 3) The mean biases, root of mean square errors (RMS), coverage probabilities of bootstrap confidence intervals (CIs) with a nominal level of 0.95 (95ECP) and CI lengths of different estimators for quantile regression coefficient with $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$, $n = 500$, and 40% missing in y_i

	$\tau = 0.1$				$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$				$\tau = 0.9$			
Method	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
$\beta_{j,\tau}^{true}$	0.507	−2.493	1.803	0.500	0.635	−2.365	1.854	0.500	0.871	−2.129	1.948	0.500	1.226	−1.774	2.090	0.500	1.664	−1.336	2.265	0.500
	Mean Biases																			
fullQr	−0.001	0.001	−0.001	0.007	−0.003	0.003	−0.002	0.010	−0.001	0.004	−0.004	0.009	0.005	0.002	−0.005	0.008	−0.009	0.003	−0.003	0.017
NaiveQr	0.095	0.076	−0.029	0.000	0.187	0.139	−0.059	−0.005	0.360	0.233	−0.116	−0.015	0.559	0.313	−0.183	−0.019	0.726	0.364	−0.240	−0.024
elmIpwQr	0.137	0.104	−0.053	0.004	0.137	0.162	−0.045	0.016	0.177	0.248	−0.043	0.039	0.387	0.415	−0.088	0.023	0.594	0.524	−0.118	−0.015
elsIpwQr	0.058	0.028	−0.036	0.019	0.061	0.047	−0.041	0.027	0.083	0.091	−0.053	0.044	0.085	0.173	−0.040	0.054	0.082	0.246	−0.030	0.094
swelQr	0.019	0.022	−0.026	0.025	0.029	0.028	−0.035	0.039	0.030	0.054	−0.043	0.064	0.003	0.131	−0.031	0.096	−0.011	0.192	−0.025	0.147
proIpwQr	0.027	0.002	−0.016	0.006	−0.009	−0.017	−0.009	0.020	−0.060	−0.037	0.011	0.023	−0.136	−0.058	0.041	0.034	−0.072	−0.035	0.054	0.045
	RMS																			
fullQr	0.113	0.057	0.059	0.089	0.161	0.077	0.081	0.124	0.231	0.114	0.117	0.183	0.357	0.178	0.186	0.286	0.603	0.302	0.304	0.478
NaiveQr	0.214	0.121	0.097	0.140	0.320	0.188	0.135	0.189	0.524	0.295	0.218	0.261	0.792	0.413	0.321	0.393	1.167	0.552	0.497	0.670
elmIpwQr	0.570	0.243	0.376	0.188	0.764	0.354	0.442	0.334	1.198	0.621	0.650	0.660	1.401	0.916	0.608	0.875	1.682	1.072	0.678	1.150
elsIpwQr	0.236	0.129	0.111	0.145	0.371	0.226	0.161	0.195	0.650	0.420	0.265	0.286	0.951	0.658	0.366	0.465	1.257	0.849	0.529	0.772
swelQr	0.210	0.120	0.098	0.130	0.344	0.212	0.147	0.178	0.620	0.392	0.247	0.270	0.934	0.642	0.367	0.469	1.180	0.770	0.504	0.766
proIpwQr	0.199	0.096	0.099	0.144	0.265	0.127	0.131	0.189	0.390	0.183	0.181	0.250	0.543	0.252	0.253	0.355	0.799	0.356	0.372	0.556
	95ECP																			
fullQr	0.948	0.947	0.938	0.946	0.943	0.955	0.948	0.959	0.936	0.957	0.951	0.961	0.948	0.950	0.943	0.945	0.940	0.938	0.947	0.952
NaiveQr	0.918	0.872	0.929	0.950	0.881	0.805	0.930	0.940	0.849	0.765	0.886	0.952	0.846	0.801	0.903	0.947	0.883	0.861	0.909	0.935
elmIpwQr	0.858	0.802	0.882	0.926	0.807	0.715	0.867	0.899	0.784	0.671	0.864	0.898	0.857	0.737	0.921	0.916	0.878	0.823	0.922	0.915
elsIpwQr	0.901	0.853	0.907	0.948	0.877	0.795	0.917	0.950	0.838	0.775	0.904	0.961	0.835	0.824	0.919	0.946	0.874	0.868	0.920	0.948
swelQr	0.895	0.910	0.908	0.939	0.898	0.904	0.934	0.957	0.880	0.910	0.938	0.961	0.877	0.918	0.911	0.939	0.833	0.882	0.860	0.886
proIpwQr	0.936	0.929	0.927	0.946	0.937	0.935	0.947	0.948	0.917	0.938	0.951	0.957	0.925	0.930	0.951	0.945	0.920	0.936	0.941	0.939
	CI lengths																			
fullQr	0.460	0.229	0.238	0.369	0.627	0.312	0.327	0.505	0.928	0.463	0.483	0.744	1.459	0.722	0.750	1.168	2.352	1.178	1.222	1.914
NaiveQr	0.785	0.380	0.375	0.559	1.061	0.508	0.509	0.751	1.553	0.722	0.742	1.095	2.341	1.087	1.133	1.660	3.650	1.683	1.745	2.620
elmIpwQr	0.885	0.441	0.470	0.659	1.289	0.614	0.663	0.924	2.127	0.963	1.027	1.395	3.248	1.485	1.532	2.137	4.517	2.139	2.153	3.199
elsIpwQr	0.747	0.362	0.368	0.557	1.071	0.519	0.522	0.782	1.596	0.786	0.781	1.157	2.399	1.207	1.176	1.779	3.482	1.722	1.727	2.658
swelQr	0.813	0.484	0.375	0.541	1.437	0.885	0.625	0.862	2.622	1.640	1.146	1.489	3.967	2.402	1.694	2.540	4.657	2.644	2.338	3.553
proIpwQr	0.802	0.384	0.406	0.572	1.094	0.531	0.538	0.770	1.534	0.745	0.740	1.066	2.130	0.998	1.039	1.507	3.005	1.415	1.488	2.230

the quantile regression model quantifies the risks of both hyperglycemia and hypoglycemia.

For illustration purposes, we limit the sample to 18744 ICU patients who are either black or white, have insurance (either government or private), and have complete records on age, gender, marital status (Married or Single), and first language (English or non-English). Among them, 7148 have glucose measurements at their admission, resulting in a 61.8% missing rate. Among the 7148 glucose measurements, 2.43% are less than 80 (hypoglycemia), and 29.25% are higher than 180 mg/dL (hyperglycemia). The observed hyperglycemia prevalence among the patients with observed glucose measures is much higher than the reported ones among ICU patients. Hyperglycemia is far more prevalent among the patients with observed glucose measures than what is reported for general ICU patient populations (eg, the proportion of hyperglycemia was 15% in Krinsley (2003) and 10% in Falciglia *et al.* (2009), and the proportion of hypoglycemia was 10% in Hulkower *et al.* (2014)). The higher hyperglycemia prevalence indicates that those with hyperglycemia are more likely to be concerned and have their glucose measured at admission, while those with normoglycemia or hypoglycemia are more likely to be underdiagnosed at admission. This indicates the possibility of missing not at random of glucose measurement.

Let Y be the glucose level at admission. We consider the following linear quantile model:

$$Q_Y(\tau \mid \mathbf{X}) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \beta_3(\tau)X_3 + \beta_4(\tau)X_4 + \beta_5(\tau)X_5 + \beta_6(\tau)X_6,$$

where X_1, X_2, X_3, X_4, and X_5 are binary indicators that the patient is female, black, reliant on government insurance, unmarried, and non-English speaking, while X_6 is the continuous variable of age.

We apply the proposed method (proIpwQr) to estimate the quantile coefficient functions with $k_n = \lfloor 2n^{0.4} \rfloor$ evenly spaced quantile grids, where $\lfloor u \rfloor$ denotes the integer part of u. For both elsIpwQr and proIpwQr, we assume the following working parametric model for the missing mechanism:

$$\text{logit}\{pr(\delta = 1 \mid \mathbf{Z}_1, Y)\} = \theta^\top \mathbf{Z}, \tag{6}$$

where $\mathbf{Z}_1 = (1, X_2, X_3, X_4, X_5, X_6)^\top$ and $\mathbf{Z} = (\mathbf{Z}_1^\top, Y)^\top$. Here, we consider the indicator variable X_1 of gender as the instrumental variable which was suggested in Shao and Wang (2016), that is, we assume that X_1 are related to the study response Y but unrelated to the propensity when the response Y and other covariates are given.

In this comparison, we could not include the method swelQr as it fails to converge numerically with this data set. We also did not include the memory-costing method elmIpwQr, since it underperforms elsIpwQr, as shown in the simulation study. Similar to the simulation study, we use the bootstrap approach developed in Section 3.2 with 100 resamples to obtain the standard errors, and we calculate the corresponding p-values from the normal approximation based on the asymptotic distributions given in Section 3.

The resulting quantile estimates, their bootstrap standard errors, and p-values are summarized in Table 2. The quantile coefficient estimates of the method elsIpwQr are very similar to those of the unadjusted method NaiveQr, which are likely to be biased. The proposed method proIpwQr, on the other hand, offered considerably different estimation and inference. For example, proIpwQr reported a significant distributional difference in glucose by race and insurance type at both lower and upper tails. Specifically, the 0.1st conditional quantile of glucose given the black population is lower than that of the white population by 7.83. Likewise, the 0.1st conditional quantile of glucose given government insurance is lower than that of the private insurance holders by 5.33. Similar differences are found at the 0.25th quantile by 3.82 and 3.51, respectively. In contrast, the 0.9th conditional quantile of glucose given black population and government insurance holders are higher by 28.00 and 7.00, respectively, compared to those of the white people and private insurance holders. Those quantile-specific effects suggest that race and insurance type are linked to the higher risk of both hypoglycemia and hyperglycemia. On the other hand, the NaiveQr and elsIpwQr reported similar patterns of distributional differences induced by race and insurance type. Still, they underestimate the lower quantiles effects and significantly overestimate the upper quantiles effects. As a result, they could have overlooked the potential risk of hypoglycemia and overestimated the risk of hyperglycemia.

The standard errors of the proposed estimates are larger than those of the NaiveQr and elsIpwQr estimates at the 0.1st quantile level but much smaller at the median or higher quantiles. That, again, indicates the right correction of quantile estimates from the proposed method. The standard error of quantile coefficients heavily depends on the data density/sparsity at the local quantile levels. They tend to be relatively small around the median where the data are dense. They then increase when the quantile level moves to either 0 or 1 and the data become locally sparse. Since the high glucose levels are oversampled in the data, the sample (observed) upper quantiles are linked to higher quantile levels when being mapped to the general population. As a result, when the proposed method estimates the standard error of a quantile estimate, it uses the local density at a lower sample quantile. Such readjusted local density is larger than the nominal density at upper quantiles and smaller at lower quantiles. This leads to larger standard

TABLE 2 The estimated coefficients (Est) and the corresponding standard errors (SE), and p-values (p) from different methods with $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$, $n = 18,744$, and 61.8% missing in response y_i

	NaiveQr			**elsIpwQr**			**proIpwQr**		
Covariate	**Est**	**SE**	***p***	**Est**	**SE**	***p***	**Est**	**SE**	***p***
	$\tau = 0.1$								
Intercept	92.07	2.34	0.00	93.04	2.30	0.00	88.02	3.80	0.00
Race	−2.18	2.09	0.30	−2.72	2.50	0.28	−7.87	2.60	0.00
Insurance type	−4.38	1.26	0.00	−4.45	1.15	0.00	−5.33	1.44	0.00
Marital status	0.67	1.07	0.53	0.55	1.15	0.63	1.30	1.56	0.40
First language	−0.83	2.79	0.77	−0.53	3.03	0.86	−1.25	5.67	0.83
Age	0.16	0.04	0.00	0.15	0.04	0.00	0.15	0.06	0.01
Gender	1.69	1.24	0.17	2.25	1.20	0.06	0.30	1.67	0.86
	$\tau = 0.25$								
Intercept	109.32	2.99	0.00	108.41	3.36	0.00	99.06	2.81	0.00
Race	0.07	2.20	0.98	−0.08	2.14	0.97	−3.82	2.62	0.15
Insurance type	−2.46	1.43	0.09	−2.36	1.36	0.08	−3.51	1.35	0.01
Marital status	1.07	1.19	0.37	0.70	1.35	0.60	−0.54	1.50	0.72
First language	4.25	4.96	0.39	4.31	5.11	0.40	−2.40	3.30	0.47
Age	0.15	0.05	0.00	0.16	0.05	0.00	0.21	0.04	0.00
Gender	4.66	1.24	0.00	5.03	1.23	0.00	5.00	1.21	0.00
	$\tau = 0.5$								
Intercept	138.92	3.74	0.00	137.62	4.31	0.00	125.69	2.84	0.00
Race	10.33	3.59	0.00	10.62	4.39	0.02	−0.05	2.66	0.99
Insurance type	−0.55	1.95	0.78	−0.87	1.90	0.65	−1.45	1.74	0.41
Marital status	1.27	1.72	0.46	1.55	1.71	0.37	0.40	1.54	0.79
First language	10.94	4.54	0.02	11.24	4.39	0.01	2.55	3.93	0.52
Age	0.11	0.07	0.10	0.13	0.07	0.08	0.18	0.05	0.00
Gender	5.28	1.66	0.00	5.70	1.57	0.00	4.29	1.51	0.00
	$\tau = 0.75$								
Intercept	191.98	7.03	0.00	189.57	6.63	0.00	151.66	5.06	0.00
Race	27.00	4.74	0.00	27.28	5.27	0.00	15.23	3.34	0.00
Insurance type	5.87	3.22	0.07	7.82	3.17	0.01	−0.69	2.43	0.78
Marital status	6.64	2.86	0.02	6.36	2.68	0.02	0.51	1.81	0.78
First language	12.18	4.67	0.01	11.17	4.74	0.02	3.05	4.56	0.50
Age	−0.21	0.12	0.07	−0.20	0.11	0.08	0.23	0.08	0.01
Gender	6.15	2.53	0.02	6.51	2.74	0.02	8.55	1.84	0.00
	$\tau = 0.9$								
Intercept	295.01	19.27	0.00	297.42	18.73	0.00	204.00	8.63	0.00
Race	62.98	15.68	0.00	62.89	17.65	0.00	28.00	6.38	0.00
Insurance type	10.13	6.25	0.11	11.03	7.99	0.17	7.00	3.88	0.07
Marital status	17.79	4.76	0.00	18.10	4.86	0.00	2.00	3.64	0.58
First language	10.24	14.39	0.48	9.56	12.84	0.46	4.00	7.33	0.59
Age	−0.96	0.28	0.00	−1.00	0.27	0.00	0.00	0.14	1.00
Gender	5.87	5.24	0.26	5.96	5.85	0.31	8.00	3.38	0.02

error estimations at lower quantiles and smaller standard error estimations at upper quantiles. Since elsIpwQr fails to correct the quantile estimates, it again produces standard error estimations comparable to NaiveQr.

5.2 | Analysis for validation

To further assess and validate the performance of the proposed method as well as the methods NaiveQr and elsIpwQr in dealing with the nonignorable missing data, we take those 7148 ICU completely observed data as a full data set. Here, we consider a more common missing case, where physicians are more likely to measure glucose of those with low or high glucose levels. Let $\tilde{Y} = (Y - \bar{Y})/\sigma(Y)$, where $\bar{Y}$ and $\sigma(Y)$ are the sample mean and standard error of the 7148 glucose measurements, respectively. We artificially create nonignorable missing responses from the missing mechanism model, $\text{logit}\{pr(\delta = 1 \mid \tilde{Y})\} = \theta_0 + \theta_1 \tilde{Y}^2$, where $\delta = 1$ if Y is observed and $\delta = 0$ if Y is missing. Here, we consider $(\theta_0, \theta_1) = (-0.6, 2.7)$ and $(-1.7, 2.7)$, which leads to the corresponding missing rates of the response Y as 40% and 60%, respectively. It is clear that the smaller $\tilde{Y}^2$ implies larger missing probability, indicating both high and low glucose levels are oversampled. For the methods elsIpwQr and proIpwQr, we use a working parametric model of the same form as model (14) with Y replaced by $\tilde{Y}^2$. In this experiment, the estimation using the 7148 ICU completely observed data (denoted as fullQr) is the benchmark. We expect the NaiveQr to be biased, and we examine how well the elsIpwQr and proIpwQr bring the estimates closer to the benchmark estimates.

Using the same estimation and bootstrapping procedures in the proceeding section, we first conduct the validation analysis once at each missing rate and present the corresponding vertical forest plots in Figure 1 to compare the estimated quantile coefficients $\hat{\boldsymbol{\beta}}_\tau$ from these different methods at the quantiles 0.1 and 0.5 and missing rate 60%. Similar results from other quantile levels and missing rates are summarized in Web Appendix F.1. In these forest plots, the ordinate lists the covariate variables; the points with different shapes represent the quantile regression coefficient estimates from different methods, while the horizontal lines represent 95% CIs. The black vertical line marks the location of zero (the figure appears in color in the electronic version of this article, and any mention of color refers to that version). If a CI crosses the zero vertical line, the corresponding coefficient is nonsignificant; otherwise, it is significant at the level of 0.05. We include the estimated intercepts and a close-up plot of Age coefficients in Figure 1B and D. Additional numerical results are summarized in Web Appendix F.2 and F.3 for readers' interest.

TABLE 3 The average coverage probabilities (covP) and mean lengths (ML) of 90% prediction intervals for glucose measurements in test set at different missing levels

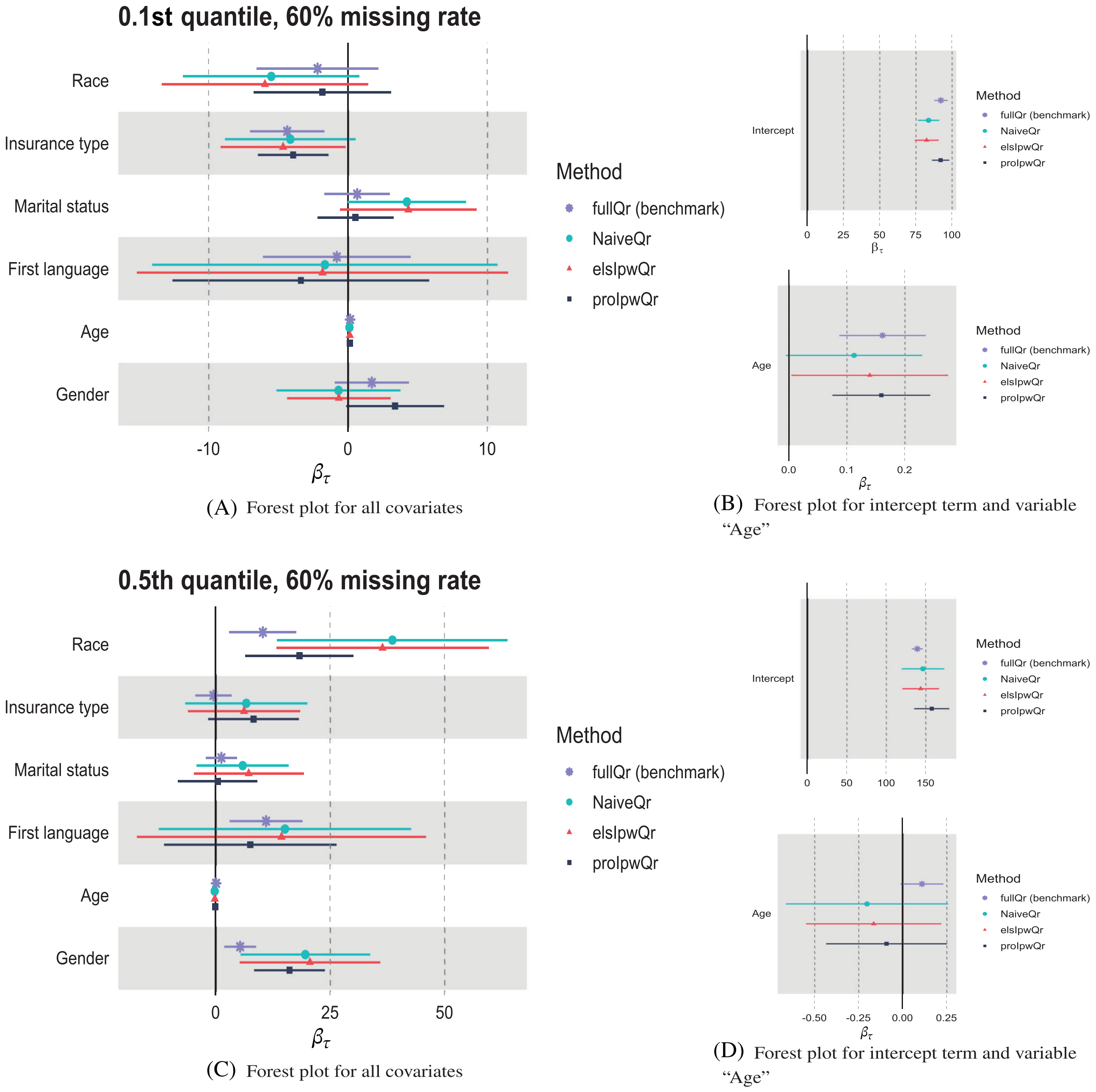
	40% Missing rate		60% Missing rate	
Method	**covP**	**ML**	**covP**	**ML**
fullQr	0.91	225.35	0.90	221.75
NaiveQr	0.94	277.48	0.96	315.81
elsIpwQr	0.94	278.49	0.96	315.83
proIpwQr	0.91	227.06	0.90	224.50

As indicated in Figure 1, the proposed method produces consistent estimates and inferences with the benchmark method fullQr, in most cases. The standard errors of the proposed estimates are smaller than those of its competitors at the 0.1st and 0.9th quantiles for both missing rates, and they are smaller or comparable at the other quantiles, which are expected by the design of the missing mechanism. The performance of the methods NaiveQr and elsIpwQr are similar, which is consistent with the results reported in Table 2, and these two methods produce very different estimates for intercept and race at the 0.75th and 0.9th quantiles from those produced from the fullQr, implying that both methods significantly overestimate the upper quantiles effects.

To further assess the prediction accuracy of these methods, we randomly split the 7148 ICU completely observed data into a training set and a test set with a ratio of 8:2. Based on the training set, we first artificially create nonignorable missing responses and then estimate the 0.05th and 0.95th quantiles regression coefficients. We use these quantile coefficient estimates to obtain the 0.05th and 0.95th conditional quantiles of the responses in the test set, and then we calculate the coverage probability and the corresponding mean length of those 90% prediction intervals for 40% and 60% missing rates. We repeat this procedure five times and report the resulting averages of the coverage probability and the mean length in Table 3. As shown in this table, the performance of the proposed method is quite close to the benchmark method fullQr, which again validates the proposed method with informative missing. Since the NaiveQr and elsIpwQr overestimate the lower and upper quantiles effects, their average coverage probability is significantly higher than 0.9, and their corresponding mean length is also much larger than that of the fullQr and proIpwQr.

6 | CONCLUSION

In this paper, we considered quantile regression with nonignorable missing responses. We proposed a joint estimating equations approach to solving the quantile and



(A) Forest plot for all covariates

(B) Forest plot for intercept term and variable "Age"

(C) Forest plot for all covariates

(D) Forest plot for intercept term and variable "Age"

FIGURE 1 The forest plots for quantile regression coefficient estimates at 0.1st and 0.5th quantiles and 60%60% missing rate. The points with different shapes represent the quantile regression coefficient estimates from different methods. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

missing mechanism models iteratively and simultaneously. We used the conditional quantile process defined from the quantile model to derive the conditional distribution of Y given $\mathbf{X}$, which facilitated the unbiased estimation of the missing data model. The implicit but deterministic relationship between the quantile model and conditional density of Y given $\mathbf{X}$ has been overlooked in the existing literature, which leads to unnecessary parametric assumptions and consequently increases the risk of biased quantile estimation. The convenient generation of random numbers from the quantile function allows us to use the Monte Carlo average to approximate the estimating equations of $\boldsymbol{\theta}$, making the estimating algorithm computationally straightforward and efficient. We have also theoretically validated the proposed estimating and its coupling bootstrap approach. As shown in both simulations and real-world examples, our method largely reduces the estimation bias and significantly improves its efficiency.

To further improve the performance of the estimation in cases with heavily missing rates, we *double-robust* the proposed estimator as in Robins *et al.* (1994), although theoretical properties and inference tools need to be reinvestigated. Another possible extension is to relax the parametric model assumption on the missing mechanism and consider semiparametric methods, as suggested by Kim and Yu (2011) and Shao and Wang (2016) to reduce the influence of misspecifications. We will leave these as future research topics.

ACKNOWLEDGMENTS

Dr. Feng's work was partially supported by the National Natural Science Foundation of China (11971292 and 11690012). Dr. Wei's work was partially supported by the NIH (R01-HG008980 R01-1RF1AG072272) and the NSF (DMS-1953527). Dr. Zhong's work was partially supported by the National Natural Science Foundation of China (11901376).

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper were derived from the Medical Information Mart for Intensive Care (MIMIC-III) database at https://physionet.org/content/mimiciii/. The data are not publicly available due to privacy or ethical restrictions. Interested researchers can apply for free access by submitting an application to PhysioNet at https://mimic.mit.edu/docs/gettingstarted/.

ORCID

Yujie Zhong https://orcid.org/0000-0002-4448-2932
Xingdong Feng https://orcid.org/0000-0002-6091-8169

REFERENCES

Aerts, M., Claeskens, G., Hens, N. & Molenberghs, G. (2002) Local multiple imputation. *Biometrika*, 89, 375–388.

Baker, L., Maley, J.H., Arévalo, A., DeMichele, F., Mateo-Collado, R., Finkelstein, S. et al. (2020) Real-world characterization of blood glucose control and insulin use in the intensive care unit. *Scientific Reports*, 10, 1–10.

Bind, M.A., Peters, A., Koutrakis, P., Coull, B. & Schwartz, J. (2015) What are the distributional distortions of air pollution on biomarkers of cardiovascular disease? *ISEE Conference Abstracts*, 2015, 1531.

Bind, M.A., Peters, A., Koutrakis, P., Coull, B., Vokonas, P. & Schwartz, J. (2016) Quantile regression analysis of the distributional effects of air pollution on blood pressure, heart rate variability, blood lipids, and biomarkers of inflammation in elderly American men: the normative aging study. *Environmental Health Perspectives*, 124, 1189–1198.

Chen, X., Wan, A.T. & Zhou, Y. (2015) Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association*, 110, 723–741.

Cheng, P.E. (1994) Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89, 81–87.

Doshi-Velez, F., Ge, Y. & Kohane, I. (2014) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133, e54–e63.

Falciglia, M., Freyberg, R.W., Almenoff, P.L., D'Alessio, D.A. & Render, M.L. (2009) Hyperglycemia-related mortality in critically ill patients varies with admission diagnosis. *Critical Care Medicine*, 37, 3001–3009.

Ghasemzadeh, S., Ganjali, M. & Baghfalaki, T. (2018) Bayesian quantile regression for analyzing ordinal longitudinal responses in the presence of non-ignorable missingness. *Metron*, 76, 321–348.

Gong, J., Lu, Y. & Xie, H. (2020) The average and distributional effects of teenage adversity on long-term health. *Journal of Health Economics*, 71, 102288.

Greenlees, J.S., Reece, W.S. & Zieschang, K.D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251–261.

Han, P. (2018) Calibration and multiple robustness when data are missing not at random. *Statistica Sinica*, 28, 1725–1740.

He, X. & Shao, Q. (2000) On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73, 120–135.

Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

Hulkower, R.D., Pollack, R.M. & Zonszein, J. (2014) Understanding hypoglycemia in hospitalized patients. *Diabetes Management*, 4, 165–176.

Inzucchi, S.E. (2006) Management of hyperglycemia in the hospital setting. *New England Journal of Medicine*, 355, 1903–1911.

Johnson, A., Pollard, T. & Mark, R. (2016a) MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.

Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M. et al. (2016b) Mimic-III, a freely accessible critical care database. *Scientific Data*, 3, 1–9.

Kim, J.K. & Riddles, M.K. (2012) Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38, 157–165.

Kim, J.K. & Yu, C.L. (2011) A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106, 157–165.

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

Koenker, R. & Bassett, G., Jr, (1978) Regression quantiles. *Econometrica*, 46(1), 33–50.

Krinsley, J.S. (2003) Association between hyperglycemia and increased hospital mortality in a heterogeneous population of critically ill patients. In: *Mayo Clinic Proceedings*, Volume 78. Elsevier, pp. 1471–1478.

Krinsley, J.S. (2004) Effect of an intensive glucose management protocol on the mortality of critically ill adult patients. In: *Mayo Clinic Proceedings*, Volume 79. Elsevier, pp. 992–1000.

Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R. et al. (2015) Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, 7, 311ra174–311ra174.

Lipsitz, S.R., Zhao, L.P. & Molenberghs, G. (1998) A semiparametric method of multiple imputation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 127–144.

Miao, W. & Tchetgen, E.T. (2018) Identification and inference with nonignorable missing covariate data. *Statistica Sinica*, 28, 2049–2067.

Miotto, R., Li, L., Kidd, B.A. & Dudley, J.T. (2016) Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 1–10.

Qin, J., Shao, J. & Zhang, B. (2008) Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, 103, 797–810.

Rao, C.R. & Zhao, L. (1992) Approximation to the distribution of m-estimates in linear models by randomly weighted bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A*, 54(3), 323–331.

Riddles, M.K., Kim, J.K. & Im, J. (2016) A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215–245.

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581–592.

Rubin, D.B. (1978) Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. In: *Proceedings of the Survey Research Methods*, Volume 1. American Statistical Association, pp. 20–34.

Shao, J. & Wang, L. (2016) Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103, 175–187.

Sherwood, B., Wang, L. & Zhou, X.H. (2013) Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in Medicine*, 32, 4967–4979.

Tang, G., Little, R.J. & Raghunathan, T.E. (2003) Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90, 747–764.

Tatonetti, N.P., Patrick, P.Y., Daneshjou, R. & Altman, R.B. (2012) Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4, 125ra31–125ra31.

van der Laan, M.J. & McKeague, I.W. (1998) Efficient estimation from right-censored data when failure indicators are missing at random. *Annals of Statistics*, 26(1), 164–182.

Wang, H.J. & Feng, X. (2012) Multiple imputation for m-regression with censored covariates. *Journal of the American Statistical Association*, 107, 194–204.

Wang, H.J., Feng, X. & Dong, C. (2019) Copula-based quantile regression for longitudinal data. *Statistica Sinica*, 29, 245–264.

Wang, S., Shao, J. & Kim, J.K. (2014) An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24, 1097–1116.

Wei, Y. & Carroll, R.J. (2009) Quantile regression with measurement error. *Journal of the American Statistical Association*, 104, 1129–1143.

Wei, Y. & Yang, Y. (2014) Quantile regression with covariates missing at random. *Statistica Sinica*, 24(3), 1277–1299.

Wu, H., Yu, X., Wang, Q., Zeng, Q., Chen, Y., Lv, J. et al. (2021) Beyond the mean: Quantile regression to differentiate the distributional effects of ambient PM2.5 constituents on sperm quality among men. *Chemosphere*, 285, 131496.

Yuan, Y. & Yin, G. (2010) Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66, 105–114.

Zhang, T. & Wang, L. (2020) Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics & Data Analysis*, 144, 106888.

Zhao, J. & Ma, Y. (2018) Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 105, 479–486.

Zhao, P., Tang, N. & Jiang, D. (2017) Efficient inverse probability weighting method for quantile regression with nonignorable missing data. *Statistics*, 51, 363–386.

Zhao, P., Tang, N., Qu, A. & Jiang, D. (2017) Semiparametric estimating equations inference with nonignorable missing data. *Statistica Sinica*, 27, 89–113.

Zhao, Z. & Xiao, Z. (2014) Efficient regressions via optimally combining quantile information. *Econometric Theory*, 30, 1272–1314.

Zou, H. & Yuan, M. (2008) Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36, 1108–1126.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2–5 are available with this paper at the Biometrics website on Wiley Online Library. An R package implementing the proposed method is available at both the Biometrics website on Wiley Online Library and https://github.com/yuaasufe/qrMNAR.

How to cite this article: Yu, A., Zhong, Y., Feng, X. & Wei, Y. (2023) Quantile regression for nonignorable missing data with its application of analyzing electronic medical records. *Biometrics*, 2036–2049. https://doi.org/10.1111/biom.13723