Statistica Sinica Preprint No: SS-2025-0156

Title	Conformal Prediction Under Nonignorable Missingness
Manuscript ID	SS-2025-0156
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0156
Complete List of Authors	Menghan Yi, Yingying Zhang, Yanlin Tang and Huixia Judy Wang
Corresponding Authors	Yanlin Tang
E-mails	yanlintang2018@163.com

Notice: Accepted author version.

Statistica Sinica

Statistica Sinica: Newly accepted Paper

Statistica Sinica

Conformal Prediction Under Nonignorable Missingness

Menghan Yi[1], Yingying Zhang[1], Yanlin Tang[1*] and Huixia Judy Wang[2*]

[1]*East China Normal University and* [2]*Rice University*

Abstract: Existing methods for handling nonignorable missing data often rely on strong modeling assumptions, making them vulnerable to model misspecification. This paper proposes a conformal prediction framework for constructing prediction sets under nonignorable missing responses, which is model-free for the outcome regression while relying on a consistently estimated propensity score. Our framework addresses two central challenges posed by nonignorable missingness: non-identifiability and the lack of data exchangeability. The key idea is to construct the highest conditional density prediction set using a local subset near the target point, while correcting for selection bias via modeling the missingness mechanism. Within this framework, we develop a bias-adjusted semiparametric method for conditional density estimation, which fits a quantile process to the observed data and corrects for bias using propensity weights. This estimator integrates seamlessly into the conformal framework, allowing our approach to guarantee not only marginal coverage, but also local and asymptotic conditional coverage for any new subject, while achieving asymptotically optimal interval lengths. We demonstrate the validity and efficiency of our procedure through simulation studies and an application to a real HIV-CD4 dataset.

Key words and phrases: Conditional density estimation; Distribution shift; Inverse probability weighting; Localized prediction; Nonignorable missing.

1. Introduction

Quantifying predictive uncertainty is a critical task in statistical modeling, especially in applications that demand reliable decision support. In practice, however, missing responses are common in both experimental and observational studies, which greatly complicates the problem. Such missingness often arises from subject dropout, unavailable measurements, or data loss, and these factors are frequently related to the response variable itself. For example, in the ACTG 175 study (Hammer et al., 1996), patients with declining CD4 cell counts were more likely to miss follow-up visits, suggesting that nonresponse may be informative (Hogan and Laird, 1997; Yuan and Yin, 2010). This type of missingness is referred to as nonignorable missingness or missing not at random (MNAR; Little and Rubin, 2019). Since the likelihood of missing data depends on the latent response, it leads to data imbalance and parameter non-identifiability, which pose significant challenges for prediction and uncertainty quantification. In this paper, we address these challenges by proposing a framework for constructing prediction sets for a new subject's response Y, conditional on covariates

Statistica Sinica

$\mathbf{X}$, when the training data contain nonignorable missingness.

Although there is a rich body of work on parameter estimation and confidence interval construction under nonignorable missingness (e.g., Zhao and Shao, 2015; Zhao and Ma, 2022; Li et al., 2022, 2023; Tian et al., 2025), predictive inference in this setting remains relatively underexplored. Existing methods typically rely on correctly specifying the regression model $Y \mid \mathbf{X}$, which is challenging in the presence of missing responses. To alleviate the impact of model misspecification, Zhao et al. (2020) proposed a generalized empirical likelihood method that does not specify the outcome model but assumes a fully parametric missing data mechanism; Miao et al. (2024) and Sun et al. (2026) proposed semiparametric methods that are doubly robust with respect to the correct specification of either the missing mechanism or the outcome regression; Li et al. (2023) proposed a fully nonparametric approach that avoids modeling both the missingness mechanism and the outcome regression by leveraging instrumental variables and imposing additional structural conditions for identification. However, these methods focus on asymptotic inference for fixed parameters or functionals and do not directly support predictive inference for a random outcome. In our work, we directly construct prediction sets for a new, unseen response. The method is model-agnostic with respect to the outcome regression $Y \mid \mathbf{X}$,

Statistica Sinica

Statistica Sinica: Newly accepted Paper

while relying on a correctly specified parametric or semiparametric model for the missingness mechanism to address non-identifiability.

Our proposed method builds on the framework of conformal inference (Vovk et al., 2005; Shafer and Vovk, 2008; Lei et al., 2018), which offers a flexible and model-agnostic approach for constructing prediction sets with finite-sample coverage guarantees. A key strength of conformal prediction is its ability to accommodate any predictive model, including black-box methods, provided that the data are exchangeable. However, this crucial assumption is violated under nonignorable missingness, where selection bias makes the observed data no longer exchangeable with the full population, thereby rendering standard conformal prediction invalid. Several studies have explored conformal prediction under non-exchangeable data. For example, Tibshirani et al. (2019) introduced a weighted conformal prediction framework to handle covariate distribution shifts between training and test data. This idea has been extended to several domains, including causal inference (Lei and Candès, 2021; Jin et al., 2023; Yin et al., 2024), survival analysis (Candès et al., 2023; Gui et al., 2024), and policy evaluation (Zhang et al., 2023). While this line of work informs our thinking, directly applying weighted conformal prediction in our setting is challenging. Unlike previous methods that only adjust for selection bias based on fully observed

$\mathbf{X}$, our framework must additionally address systematic bias introduced by partially observed Y, which gives rise to non-identifiability issues.

This paper addresses the challenges of non-exchangeability and non-identifiability arising from nonignorable missingness, with key innovations and contributions highlighted in three main aspects. First, we introduce a novel MNAR-weighted conformal prediction framework for predicting outcomes in new subjects, where the training data exhibit nonignorable missingness. This framework quantifies and corrects the selection bias and non-exchangeability induced by missingness via density ratio weighting, while flexibly incorporating parametric or semiparametric models for the missingness mechanism to tackle identifiability challenges. Unlike conventional methods (Tibshirani et al., 2019; Lei and Candès, 2021) that account only for covariate shift, our framework provides a more comprehensive correction for selection bias that arises from the joint distribution of $\mathbf{X}$ and Y.

Second, to achieve conditional coverage, we target the highest conditional density region of Y given $\mathbf{x}$, while using the profile distance to partition the covariate space and identify a local subset whose conditional density profiles are similar to that of the target point $\mathbf{x}$. This approach naturally adapts the prediction set to the local structure of the data, allowing for personalized and efficient inference. As a result, we establish theoretical

Statistica Sinica: Newly accepted Paper

guarantees for asymptotically optimal prediction sets that achieve local coverage and approximate conditional coverage for individual subjects, going beyond the marginal guarantees provided by standard conformal methods (Vovk et al., 2005; Tibshirani et al., 2019). This enables personalized adaptation to subject-level heterogeneity, even under complex error distributions such as asymmetric or multi-modal ones.

Third, we propose a bias-adjusted semiparametric procedure to estimate the conditional density, which is needed both to identify the highest predictive density region and to select a local subset. We first approximate it using quantile regression on the observed data and then calibrate it using an adjustment factor that captures the influence of Y on the missingness mechanism, allowing us to recover the conditional distribution up to a multiplicative factor. The procedure is practical and effective, and integrates seamlessly with the conformal prediction framework to construct valid and efficient prediction sets that achieve the desired conditional coverage.

The remainder of the paper is organized as follows. In Section 2, we formally present two proposed methods: non-local and localized conformal prediction. Section 3 establishes their theoretical guarantees. The performance of the proposed method is assessed through simulation studies in Section 4, and the analysis of an AIDS clinical trial dataset in Section 5. The

Statistica Sinica: Newly accepted Paper

online Supplementary Materials contain technical proofs, high-dimensional extensions, and additional numerical experiments.

2. Proposed Method

2.1 Setup and Motivation

Let $Y \in \mathcal{Y}$ denote the univariate response of interest and $\mathbf{X} \in \mathcal{X}$ represent the p-dimensional covariates, where $\mathbf{X}$ is fully observed, but Y is subject to nonignorable missingness. Define δ as the missingness indicator for Y, where $\delta = 1$ if Y is observed and $\delta = 0$ otherwise. Nonignorable missingness implies that even after conditioning on $\mathbf{X}$, the propensity score $\mathbb{P}(\delta = 1|\mathbf{X}, Y)$ still depends on the potentially missing Y. Let $F(\mathbf{x}, y, \delta)$ denote the joint distribution of the latent variables $(\mathbf{X}, Y, \delta)$, and let $\{(\mathbf{X}_i, Y_i, \delta_i) : i = 1, \ldots, n\}$ be independent and identically distributed draws from $F(\mathbf{x}, y, \delta)$. Given a new subject $(\mathbf{X}_{n+1}, Y_{n+1}) \sim F(\mathbf{x}, y)$, our goal is to predict the unknown response Y_{n+1}, based on the observed data $\{(\mathbf{X}_i, \delta_i Y_i, \delta_i) : i = 1, \ldots, n\}$ and $\mathbf{X}_{n+1}$. Denote the prediction set as $\widehat{C}(\mathbf{X}_{n+1}; \alpha)$, a subset of $\mathcal{Y}$, which satisfies the coverage guarantee,

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1}; \alpha)\} \geq 1 - \alpha, \tag{2.1}$$

Statistica Sinica: Newly accepted Paper

for a given miscoverage level $\alpha \in (0,1)$, where the probability $\mathbb{P}$ is taken over all the randomness in the data.

If the data is fully observed, we can directly use standard conformal prediction methods Vovk et al. (2005); Shafer and Vovk (2008); Lei et al. (2018) to construct prediction sets that ensure marginal validity (2.1). However, when the response variable is missing, selection bias in the observed data causes a distribution shift from the test data,

$$\{(\mathbf{X}_i, Y_i) : \delta_i = 1\} \sim F(\mathbf{x}, y \mid \delta = 1) \quad \text{and} \quad (\mathbf{X}_{n+1}, Y_{n+1}) \sim F(\mathbf{x}, y), \qquad (2.2)$$

posing a significant challenge as it violates data exchangeability and renders standard conformal prediction invalid. To address this challenge, we introduce a weighted correction to conformal prediction that adjusts for nonexchangeability by accounting for the joint effect of $\mathbf{X}$ and Y in the missingness mechanism.

Moreover, our goal is not just to satisfy the coverage lower bound (2.1) but to achieve more efficient predictions. To this end, we identify the highest conditional density set of Y as the prediction set, ensuring that the length is asymptotically optimal for any error distribution. However, conditional density estimation is particularly challenging in cases of non-ignorable missingness, where the absence of Y results in distributional non-identifiability (Robins and Ritov, 1997; Miao et al., 2016). To address this challenge,

Statistica Sinica: Newly accepted Paper

we develop a bias-adjusted semiparametric method for conditional density estimation that fits the quantile process to observed data and incorporates weights derived from the missingness propensity model to correct for bias.

2.2 MNAR-weighted Conformal Prediction

The core idea of our method is to select all candidate values $y \in \mathcal{Y}$ that are consistent with the model trained on the observed data. To evaluate this consistency, we define a nonconformity score $R(\mathbf{X}_{n+1}, y)$ that measures how much y deviates from the model — higher scores indicate greater deviation. Then, we include all candidates in the prediction set if their nonconformity scores are smaller than a certain threshold, which is chosen as the estimated $(1-\alpha)$-quantile of $R\left(\mathbf{X}_{n+1}, Y_{n+1}\right)$ to ensure $(1-\alpha)$ coverage (2.1).

Specifically, we randomly split the observed data into two parts: a training set $\mathcal{D}_t$ for constructing the nonconformity score function and a calibration set $\mathcal{D}_c$ for determining the threshold. The nonconformity score function is defined as a mapping from a data point to a real number, $R(\mathbf{x}, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures how well $(\mathbf{x}, y)$ aligns with the model trained on $\mathcal{D}_t$. Although the nonconformity score can be chosen flexibly without affecting the coverage guarantee, its choice significantly influences the size of the prediction set. To ensure prediction efficiency, we use the

Statistica Sinica: Newly accepted Paper

negative estimated conditional density as the nonconformity score,

$$R(\mathbf{x}, y) = -\widehat{f}(y \mid \mathbf{x}), \tag{2.3}$$

where $\widehat{f}$ denotes the estimated conditional density of Y given $\mathbf{x}$. Under this scoring rule, higher density indicates better consistency between the data point $(\mathbf{x}, y)$ and the estimated model $\widehat{f}$. Compared to standard residual-based methods (Shafer and Vovk, 2008; Lei et al., 2018), our experience shows that the conditional density-based approach can yield shorter prediction intervals, especially under complex error distributions such as skewed or bimodal ones. However, estimating the conditional density is challenging under nonignorable missingness. We propose an effective estimator that uses only fully observed data, with a suitable correction informed by the missing mechanism; see Section 2.4 for details.

To determine the threshold, we need an estimate of the $(1-\alpha)$-quantile of $R(\mathbf{X}_{n+1}, Y_{n+1})$. Under exchangeability, this quantile can be obtained from the empirical distribution of nonconformity scores on the calibration set $\mathcal{D}_c$. However, when missingness is present, the distribution shift (2.2) between the observed data and the test point makes this empirical quantile unreliable, and prediction coverage is no longer guaranteed. To overcome

Statistica Sinica: Newly accepted Paper

this issue, we derive the distribution function of $R(\mathbf{X}_{n+1}, Y_{n+1})$ as:

$$\begin{aligned}&\mathbb{P}_{(\mathbf{X}_{n+1},Y_{n+1})\sim F(\mathbf{x},y)}\{R(\mathbf{X}_{n+1}, Y_{n+1}) \leq r\}\\ &\quad= \mathbb{E}_{(\mathbf{X},Y)\sim F(\mathbf{x},y|\delta=1)}\left[w(\mathbf{X}, Y)\,\mathbb{I}\{R(\mathbf{X}, Y) \leq r\}\right], \qquad (2.4)\end{aligned}$$

where $w(\mathbf{X}, Y) = f(\mathbf{X}, Y)/f(\mathbf{X}, Y \mid \delta = 1)$ is referred to as a density-ratio weight function. This form of weighting is closely related to the covariate shift weights in (Tibshirani et al., 2019), but allowing for a more general setting. We estimate the weights using the training set $\mathcal{D}_t$, and the corresponding estimation procedure together with the associated identifiability issues are discussed in Section 2.5. Given the distribution in (2.4), we approximate it by constructing a weighted empirical distribution function based on the observed data. Specifically, let the observed calibration set be $\mathcal{D}_c^{\text{obs}} = \{i \in \mathcal{D}_c : \delta_i = 1\}$. Then, using the augmented set $\mathcal{D}_c^{\text{obs}} \cup \{(\mathbf{X}_{n+1}, y)\}$, with y as a candidate value for Y_{n+1}, we construct the corresponding empirical distribution function:

$$\sum_{i\in\mathcal{D}_c^{\text{obs}}} \varpi_i(y)\,\mathbb{I}\{R(\mathbf{X}_i, Y_i) \leq r\} + \varpi_{n+1}(y)\,\mathbb{I}\{R(\mathbf{X}_{n+1}, y) \leq r\}, \qquad (2.5)$$

where the normalized weight is given by:

$$\varpi_i(y) = \frac{w(\mathbf{X}_i, Y_i)}{\sum_{j\in\mathcal{D}_c^{\text{obs}}} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)}, \quad i \in \mathcal{D}_c^{\text{obs}},$$

$$\varpi_{n+1}(y) = \frac{w(\mathbf{X}_{n+1}, y)}{\sum_{j\in\mathcal{D}_c^{\text{obs}}} w(\mathbf{X}_j, Y_j) + w(\mathbf{X}_{n+1}, y)}.$$

Statistica Sinica: Newly accepted Paper

Therefore, the $(1-\alpha)$ quantile from the empirical distribution (2.5) is the desired threshold, denoted as $\widehat{r}_{\alpha}(\mathbf{X}_{n+1}, y)$. The prediction set can be constructed as the set of all y such that $R(\mathbf{X}_{n+1}, y) \leq \widehat{r}_{\alpha}(\mathbf{X}_{n+1}, y)$. We summarize the overall procedure in Algorithm 1.

Algorithm 1: MNAR-weighted Conformal Prediction

```
Input: Dataset $\{(\mathbf{X}_i, \delta_i Y_i, \delta_i) : i = 1, \dots, n\}$, test point $\mathbf{X}_{n+1}$,
       candidate grid $\mathcal{Y}_{\text{grid}} = \{y_1, y_2, \dots\} \subseteq \mathcal{Y}$, target level $\alpha \in (0, 1)$.
Split the data into two equal-sized subsets $\mathcal{D}_t$ and $\mathcal{D}_c$.
Use $\mathcal{D}_t$ to fit the nonconformity score $R(\mathbf{x}, y)$ and the weight $w(\mathbf{x}, y)$.
for $y \in \mathcal{Y}_{\text{grid}}$ do
    Use $\mathcal{D}_c \cup (\mathbf{X}_{n+1}, y)$ to compute the threshold $\widehat{r}_{\alpha}(\mathbf{X}_{n+1}, y)$, i.e., the
      $(1-\alpha)$ quantile of (2.5).
Output: Return the $(1-\alpha)$ prediction set
        $\widehat{C}(\mathbf{X}_{n+1}; \alpha) = \{y : R(\mathbf{X}_{n+1}, y) \leq \widehat{r}_{\alpha}(\mathbf{X}_{n+1}, y)\}$.
```

Remark 1 (Numerical Implementation)**.** Our conformal prediction set is defined by inverting a hypothesis test over the response space. When the response is continuous, the resulting prediction region is a (possibly disconnected) subset of $\mathbb{R}$. In practice, we obtain this region by evaluating the fitted nonconformity score on a fine grid $\mathcal{Y}_{\text{grid}} = \{y_1, y_2, \dots\} \subseteq \mathcal{Y}$, which provides a numerical approximation to the boundary of the conformal set. The grid is typically chosen as a set of uniformly spaced points with a prespecified resolution over an empirical range (Chen et al., 2018; Lei, 2019). Our method follows the split conformal prediction paradigm (Lei et al.,

Statistica Sinica

Statistica Sinica: Newly accepted Paper

2018), in which the nonconformity score and the weight function are fitted once on $\mathcal{D}_t$ and then held fixed. As a result, the grid search involves only forward evaluations of these two fitted functions, allowing the grid resolution to be increased to improve numerical accuracy at small additional computational cost.

2.3 Localized Prediction

The method introduced in Section 2.2 achieves valid marginal coverage (see Theorem 1). However, in many applications, it is desirable to provide personalized predictions by ensuring valid coverage guarantees for each new individual $\mathbf{x}_{n+1}$. To this end, we propose an enhanced approach—localized prediction—that partitions the feature space and identifies a local calibration dataset within the same cluster as $\mathbf{x}_{n+1}$ for constructing the prediction, thereby reducing the discrepancy between calibration and $\mathbf{x}_{n+1}$ to better approximate conditional coverage.

Specifically, we use K-means clustering to partition the covariate space, based on the profile distance proposed by Izbicki et al. (2022), defined as

$$d^2(\mathbf{x}_1, \mathbf{x}_2) := \int_0^\infty \left\{G_f(t \mid \mathbf{x}_1) - G_f(t \mid \mathbf{x}_2)\right\}^2 dt, \tag{2.6}$$

Statistica Sinica: Newly accepted Paper

where $G_f(t \mid \mathbf{x})$ denote the conditional CDF of $f(Y|\mathbf{X})$, given by

$$G_f(t \mid \mathbf{x}) = \mathbb{P}(f(Y \mid \mathbf{X}) \leq t \mid \mathbf{X} = \mathbf{x}) = \int_{\{y : f(y|\mathbf{x}) \leq t\}} f(y \mid \mathbf{x}) \mathrm{d}y, \tag{2.7}$$

and can be estimated by $\widehat{G}_{\widehat{f}}(t \mid \mathbf{x}) = \int_{\{y : \widehat{f}(y|\mathbf{x}) \leq t\}} \widehat{f}(y \mid \mathbf{x}) dy$. The profile distance measures the overall similarity between the profiles of the conditional densities $f(Y \mid \mathbf{X})$. It not only aligns well with our nonconformity score but also performs effectively in high-dimensional settings. Let $A_n(\mathbf{X}_{n+1})$ denote the local subset of the calibration $\mathcal{D}_c^{\text{obs}}$ that belongs to the same cluster as $\mathbf{X}_{n+1}$. The prediction set is then obtained by constructing an empirical distribution similar to (2.5), with $\mathcal{D}_c^{\text{obs}}$ replaced by $A_n(\mathbf{X}_{n+1})$. We summarize the overall procedure in Algorithm 2.

```
Algorithm 2: Localized MNAR-weighted Conformal Prediction
  Input: Dataset $\{(\mathbf{X}_i, \delta_i Y_i, \delta_i) : i = 1, \ldots, n\}$, test point $\mathbf{X}_{n+1}$,
         candidate grid $\mathcal{Y}_{\text{grid}} = \{y_1, y_2, \ldots\} \subseteq \mathcal{Y}$, target level $\alpha \in (0, 1)$,
         number of clusters.
  Split the data into two equal-sized subsets $\mathcal{D}_t$ and $\mathcal{D}_c$.
  Use $\mathcal{D}_t$ to (i) fit the nonconformity score $R(\mathbf{x}, y)$ and the weight
    $w(\mathbf{x}, y)$, and (ii) perform K-means clustering in the covariate space
    based on the profile distance (2.6).
  Define $A_n(\mathbf{X}_{n+1}) \subset \mathcal{D}_c$ as the subset sharing the same cluster as $\mathbf{X}_{n+1}$.
  for $y \in \mathcal{Y}_{\text{grid}}$ do
      Compute the threshold $\widehat{r}_\alpha(A_n; \mathbf{X}_{n+1}, y)$ as the $(1 - \alpha)$ quantile of
        (2.5) with $\mathcal{D}_c^{\text{obs}}$ replaced by $A_n(\mathbf{X}_{n+1}) \cup (\mathbf{X}_{n+1}, y)$.
  Output: Return $(1 - \alpha)$ prediction set
          $\widehat{C}(\mathbf{X}_{n+1}; \alpha) = \{y : R(\mathbf{X}_{n+1}, y) \leq \widehat{r}_\alpha(A_n; \mathbf{X}_{n+1}, y)\}$.
```

Statistica Sinica: Newly accepted Paper

2.4 Conditional Density Estimation

Most existing work on nonignorable missing data focuses on parameter estimation, with limited attention paid to the entire density function. In this section, we propose a bias-adjusted semiparametric approach to estimate the conditional density (2.3) by fitting the quantile process to the fully observed data, with an appropriate correction for the missingness mechanism. In particular, we model conditional quantiles rather than the conditional density directly, since the density admits a simple and natural representation in terms of quantiles, and model diagnosis and assessment can be more conveniently conducted on conditional quantiles. This representation allows us to leverage a wide range of existing quantile regression methods to estimate the conditional density in a flexible and robust way, even in high-dimensional settings. In contrast, existing methods, such as kernel-based approaches, tend to become increasingly difficult to implement and tune as the covariate dimension increases.

Our approach is motivated by two equivalent decompositions of the joint distribution of y and $\delta = 1$ given $\mathbf{x}$, namely $\mathbb{P}(\delta = 1 \mid \mathbf{x}, y) f(y \mid \mathbf{x})$ and $f(y \mid \mathbf{x}, \delta = 1)\mathbb{P}(\delta = 1 \mid \mathbf{x})$. Therefore, we can derive the relationship between the conditional distribution of the response and that of the observed data:

$$f(y \mid \mathbf{x}) = \rho(\mathbf{x}, y) f(y \mid \mathbf{x}, \delta = 1), \tag{2.8}$$

Statistica Sinica: Newly accepted Paper

where $\rho(\mathbf{x}, y) = \mathbb{P}(\delta = 1 \mid \mathbf{x})/\mathbb{P}(\delta = 1 \mid \mathbf{x}, y)$ is the adjustment factor, capturing how the probability of being observed changes when y is further taken into account given $\mathbf{x}$. This adjustment gives more weight to data points whose observation probabilities are smaller, conditional on $\mathbf{x}$.

We propose to estimate $f(y \mid \mathbf{x})$ based on equation (2.8), by first using the fully observed data to estimate $f(y \mid \mathbf{x}, \delta = 1)$, and then correcting it with $\rho(\mathbf{x}, y)$. To estimate $f(y \mid \mathbf{x}, \delta = 1)$, we propose a semiparametric approach by fitting a quantile process based on the observed data. The key idea is that the conditional density can be related to the conditional quantile function through the following relationship:

$$f\{Q_Y(\tau \mid \mathbf{x}, \delta = 1) \mid \mathbf{x}, \delta = 1\} = \left\{\frac{\mathrm{d}Q_Y(\tau \mid \mathbf{x}, \delta = 1)}{\mathrm{d}\tau}\right\}^{-1}, \tag{2.9}$$

where $Q_Y(\tau \mid \mathbf{x}, \delta = 1)$ is the τ-th conditional quantile of Y given $\mathbf{x}$ and $\delta = 1$. Therefore, according to (2.9), we can construct quotient-type density estimates (Siddiqui, 1960), given by:

$$\begin{aligned}&\widehat{f}\{\widehat{Q}_Y(\tau \mid \mathbf{x}, \delta = 1) \mid \mathbf{x}, \delta = 1\} \\ &\qquad = \frac{2h_n}{\widehat{Q}_Y(\tau + h_n \mid \mathbf{x}, \delta = 1) - \widehat{Q}_Y(\tau - h_n \mid \mathbf{x}, \delta = 1)},\end{aligned} \tag{2.10}$$

where $\widehat{Q}_Y$ denotes the estimated conditional quantile function, and h_n is a bandwidth parameter tending to zero as $n \to \infty$. To estimate the density $\widehat{f}(y|\mathbf{x}, \delta = 1)$ for any $y \in \mathcal{Y}$, we first compute the density along the quantile

Statistica Sinica: Newly accepted Paper

process $\{\widehat{Q}_Y(\tau_1|\mathbf{x}, \delta = 1), \dots, \widehat{Q}_Y(\tau_\kappa|\mathbf{x}, \delta = 1)\}$ using (2.10), at a set of quantile levels $\tau_1 < \cdots < \tau_\kappa$. We then apply a smoothing technique, such as linear interpolation, to recover a continuous estimate of the full density function. Note that the conditional quantiles in (2.10) can be estimated using any suitable method on the observed data $\mathcal{D}_t$ without affecting the coverage guarantee, such as linear quantile regression (Koenker, 2005), high-dimensional penalized quantile regression (Belloni and Chernozhukov, 2011; Tan et al., 2022; Qiu et al., 2026), or nonparametric approaches like quantile random forests (Meinshausen, 2006; Athey et al., 2019).

To estimate the adjustment factor $\rho(\mathbf{x}, y)$, we can directly obtain an estimate of $\mathbb{P}(\delta = 1 \mid \mathbf{x}, y)$ through the weight function estimation process in Section 2.5. Additionally, the propensity $\mathbb{P}(\delta = 1 \mid \mathbf{x})$ can be easily estimated using binary classification methods, such as logistic regression or gradient boosting, based on the fully observed data $\{(\mathbf{X}_i, \delta_i)\}$. We provide further discussion on conditional density estimation in high-dimensional settings in Section S2 of the Supplementary Material.

It is important to note that we do not use the quotient-type density to directly estimate $f(y \mid \mathbf{x})$, as estimating the conditional quantile $Q_Y(\tau \mid \mathbf{x})$ is particularly challenging under nonignorable missingness in Y, and typically requires specifying parametric models or making additional assump-

Statistica Sinica: Newly accepted Paper

tions (Zhang and Wang, 2020; Yu et al., 2023).

2.5 Weight Estimation

If the weight function is known, the constructed prediction set guarantees $(1-\alpha)$ coverage in a finite sample without any additional assumptions. However, in most applications, the weight function is unknown and needs to be estimated. To this end, we first derive the weight function as follows:

$$\begin{aligned} w(\mathbf{x}, y) &= \frac{f(\mathbf{x}, y)}{f(\mathbf{x}, y|\delta=1)} = \frac{f(\mathbf{x})}{f(\mathbf{x}|\delta=1)} \cdot \frac{f(y|\mathbf{x})}{f(y|\mathbf{x}, \delta=1)} \\ &= \frac{\mathbb{P}(\delta=1)}{\mathbb{P}(\delta=1|\mathbf{x})} \cdot \frac{\mathbb{P}(\delta=1|\mathbf{x})}{\mathbb{P}(\delta=1|\mathbf{x}, y)} \propto \frac{1}{\mathbb{P}(\delta=1 \mid \mathbf{x}, y)}. \end{aligned} \tag{2.11}$$

It follows that the weight function is entirely determined by the missing propensity score $\mathbb{P}(\delta=1 \mid \mathbf{x}, y)$. If the missing mechanism is Missing at Random (MAR), the weight function depends only on $\mathbf{x}$. Therefore, it can be directly estimated based on the fully observed data $(\mathbf{x}, \delta)$. However, under non-ignorable missingness, the weight function also depends on the missing y, making its identification and estimation more challenging.

Robins and Ritov (1997) and Miao et al. (2016) pointed out that, under nonignorable missingness, identifiability remains a major challenge even when both the missingness mechanism $\mathbb{P}(\delta=1 \mid \mathbf{x}, y)$ and the outcome model $f(y \mid \mathbf{x})$ are specified. To address this issue, many studies impose parametric or semi-parametric models while adding extra assumptions. We

Statistica Sinica: Newly accepted Paper

adopt several existing options to model the missingness mechanism.

Model 1 (Parametric Model). For nonignorable missing data, a commonly used missingness mechanism is the logistic regression model:

$$\mathbb{P}(\delta = 1 \mid \mathbf{X}, Y) = \frac{1}{1 + \exp(\alpha + \mathbf{X}^\top \boldsymbol{\beta} + Y\gamma)},$$

where $(\alpha, \boldsymbol{\beta}, \gamma)$ are unknown parameters. Under this model, Liu et al. (2022) further assumes that $f(y \mid \mathbf{x}, \delta = 1)$ follows a finite-dimensional parametric model, leveraging the information from fully observed $\mathbf{X}$ to resolve the identifiability issue and estimate (α, β, γ).

Model 2 (Semiparametric Model). To mitigate the risk of model misspecification, a more flexible semiparametric model can be adopted:

$$\mathbb{P}(\delta = 1 \mid \mathbf{X}, Y) = \frac{1}{1 + g(\mathbf{U})q\left(Y, \boldsymbol{\gamma}\right)}, \tag{2.12}$$

where $q(Y, \boldsymbol{\gamma})$ is a parametric function with an l-dimensional unknown parameter $\boldsymbol{\gamma}$, $g(\cdot)$ is an unknown nonparametric function, and $\mathbf{U} \subset \mathbf{X}$ denotes a subset of covariates that the missingness mechanism depends on. Define $\mathbf{Z} = \mathbf{X} \setminus \mathbf{U}$ as the nonresponse instrumental variables, which are related to the response variable but excluded from the missingness mechanism. By appropriately selecting $\mathbf{Z}$ and leveraging them to construct additional estimating equations, the model parameters can be rendered identifiable (Shao and Wang, 2016). Similar instrumental variable methods, with notable in-

Statistica Sinica: Newly accepted Paper

novations, have been widely used (Zhang et al., 2018; Zhao et al., 2020, 2021; Li et al., 2022, 2023; Miao et al., 2024), including in high-dimensional settings (Wang et al., 2021).

In practical applications, the selection of a parametric or semiparametric model depends on the data characteristics, while the feasibility of the instrumental variable method is also taken into account. For simplicity, we focus on Model 2 (Shao and Wang, 2016) in the following theoretical derivations and numerical simulations for detailed analysis and discussion.

It is worth noting that our method is only partially model-free: it does not require a correctly specified regression model for $Y \mid \mathbf{X}$, but relies on a correctly specified or consistently estimated propensity score $\mathbb{P}(\delta = 1 \mid \mathbf{X}, Y)$, as shown in Theorems 1 and 2. Relaxing the dependence on correct specification of the missingness mechanism remains an open challenge.

3. Theoretical Results

In this section, we establish the coverage guarantee for the proposed prediction set, beginning with the following assumption.

Assumption 1. Assume that $F(\mathbf{x}, y)$ is absolutely continuous with respect to $F(\mathbf{x}, y \mid \delta = 1)$

This assumption ensures that $w(\mathbf{x}, y)$ is finite almost surely and that the empirical distribution in (2.5) is well defined, which yields the coverage lower bound in Theorem 1 under the propensity score model (2.12).

Theorem 1. *Under Assumption 1, suppose the estimated function* $\widehat{g}(\cdot)$ *and parameter* $\widehat{\boldsymbol{\gamma}}$ *satisfy* $\mathbb{E}_{(\mathbf{X},Y)\sim f(\mathbf{x},y|\delta=1)}\{\widehat{g}(\mathbf{U})q(Y,\widehat{\boldsymbol{\gamma}}) \mid \mathcal{D}_t\} < \infty$, *where* $\mathbf{U} = \mathbf{X}\backslash\mathbf{Z}$ *denotes the covariates excluding the instrumental variables. Then, for any given* $\alpha \in (0,1)$, *the proposed methods in Algorithms 1 and 2 satisfy*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1};\alpha)\} \geq 1 - \alpha - \widehat{\Delta}_{\text{marg}},$$

where the probability is over $\mathcal{D}_c$ *and* $(\mathbf{X}_{n+1}, Y_{n+1})$, *and*

$$\widehat{\Delta}_{\text{marg}} = \frac{1}{2}\mathbb{E}_{(\mathbf{X},Y)\sim f(\mathbf{x},y|\delta=1)}\left|\widehat{g}(\mathbf{U})q(Y,\widehat{\boldsymbol{\gamma}}) - g(\mathbf{U})q(Y,\boldsymbol{\gamma})\right|.$$

Theorem 1 indicates that, when the propensity score is known, the finite-sample coverage achieves the exact $(1-\alpha)$ level regardless of whether the nonconformity score is correctly specified or consistently estimated. However, when the propensity score is unknown, the estimation bias introduces an additional error term, $\widehat{\Delta}_{\text{marg}}$, for the coverage lower bound. This error term is asymptotically negligible provided that the propensity score is consistently estimated, a condition typically guaranteed under mild assumptions (Shao and Wang, 2016; Zhao et al., 2021).

Theorem 1 shows that our method achieves average coverage over the entire population of $\mathbf{X}$, but this does not imply coverage guarantees for any specific $\mathbf{x}$. To address this limitation, Theorem 2 establishes a stronger result of local coverage, showing that our localized method guarantees coverage for each region in a prespecified partition of the covariate space.

Theorem 2. *Let $\mathcal{A} = \{A_j : j \geq 1\}$ be a partition of the covariate space $\mathcal{X}$. Under the same assumptions as Theorem 1, for any given $\alpha \in (0,1)$, the localized prediction in Algorithm 2 satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1};\alpha) \mid \mathbf{X}_{n+1} \in A_j\} \geq 1 - \alpha - \widehat{\Delta}_{\text{loc}}, \quad \textit{for all } j,$$

where the probability is over $\mathcal{D}_c$ and $(\mathbf{X}_{n+1}, Y_{n+1})$, and

$$\widehat{\Delta}_{\text{loc}} = \frac{(1-\alpha/2)\mathbb{E}_{(\mathbf{X},Y)\sim f(\mathbf{x},y|\delta=1)}\left|\widehat{g}(\mathbf{U})q(Y,\widehat{\boldsymbol{\gamma}}) - g(\mathbf{U})q(Y,\boldsymbol{\gamma})\right|}{\mathbb{E}_{\mathbf{X}\sim f(\mathbf{x}|\delta=1)}w(\mathbf{X})\mathbb{I}\left(\mathbf{X}\in A_j\right)}.$$

The local validity in Theorem 2 must be interpreted in conjunction with the resolution of the partition $\mathcal{A}$: if $\mathcal{A} = \mathcal{X}$, then local validity reduces to marginal validity; if $A_j \in \mathcal{A}$ shrinks to a single point $\mathbf{x}$, then local validity approximates conditional validity. In our procedure, we use K-means clustering based on the profile distance Izbicki et al. (2022) to partition the covariate space in a data-adaptive manner. This approach effectively captures the local structure of $Y \mid \mathbf{X}$ and better approximates the neighborhood around a given $\mathbf{x}$ compared to uniform partitioning (Lei and Wasserman, 2014). The coverage error $\widehat{\Delta}_{\text{loc}}$ in Theorem 2 is typically asymptotically negligible: under standard regularity conditions, consistent

Statistica Sinica: Newly accepted Paper

propensity score estimation ensures that the numerator approaches zero, while the denominator remains positive because the data-driven partitioning usually yields regions with nonzero probability.

While Theorems 1 and 2 establish valid lower bounds on coverage, our primary goal is to achieve optimal efficiency by constructing the smallest possible prediction sets. To evaluate the performance of our method, we begin by introducing the oracle prediction set, defined as the highest prediction density set (Izbicki et al., 2022), which serves as a benchmark.

Definition 1. The highest prediction density (HPD) set is defined as

$$C(\mathbf{x};\alpha) = \{y : f(y \mid \mathbf{x}) \geq Q_f(\alpha \mid \mathbf{x})\},$$

where $Q_f(\alpha \mid \mathbf{x})$ denotes the α-th conditional quantile of the conditional density values $f(Y \mid \mathbf{x})$. This set satisfies the conditional coverage condition $\mathbb{P}\{Y \in C(\mathbf{x};\alpha) \mid \mathbf{X} = \mathbf{x}\} = 1-\alpha$ and has the smallest Lebesgue measure among all sets with conditional coverage at least $1-\alpha$.

To establish a stronger result regarding efficiency, we introduce the following additional assumptions.

Assumption 2. The weight function $w(\mathbf{X}, Y)$ and its estimator $\widehat{w}(\mathbf{X}, Y)$ are almost surely bounded away from 0 and ∞. In addition, there exist

Statistica Sinica: Newly accepted Paper

sequences $\eta_n = o(1)$ and $\rho_n = o(1)$ such that

$$\mathbb{P}\Big(\mathbb{E}[\sup_{y\in\mathcal{Y}} \big|\widehat{w}(\mathbf{X}, y) - w(\mathbf{X}, y)\big|^2 \mid \widehat{w}] \ \geq \ \eta_n\Big) \leq \rho_n.$$

Assumption 3. There exist sequences $\eta_n = o(1)$ and $\rho_n = o(1)$ such that, for each region A_j in the partition $\mathcal{A}$,

$$\mathbb{P}\Big(\sup_{\mathbf{x}\in A_j}\sup_{y\in\mathcal{Y}} \big|\widehat{f}(y \mid \mathbf{x}) - f(y \mid \mathbf{x})\big| \ \geq \ \eta_n\Big) \leq \rho_n.$$

Assumption 4. The conditional CDF $G_f(t \mid \mathbf{x})$ in (2.7) is Lipschitz continuous in $(t, \mathbf{x})$; that is, there exists a constant $L > 0$ such that, for all $t_1, t_2 \in \mathbb{R}$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$, $\big|G_f(t_1 \mid \mathbf{x}_1) - G_f(t_2 \mid \mathbf{x}_2)\big| \leq L\big(|t_1 - t_2| + d(\mathbf{x}_1, \mathbf{x}_2)\big)$.

Assumption 5. For each region A_j in the partition $\mathcal{A}$, the probability mass satisfies $n\,\mathbb{P}\{\mathbf{X} \in A_j\} \to \infty$, and the diameter $\sup_{\mathbf{x}_1,\mathbf{x}_2\in A_j} d(\mathbf{x}_1, \mathbf{x}_2)$ converges to zero as the partition is refined.

Assumptions 2 and 3 are not difficult to satisfy, as they only require uniform consistency of the propensity score and the conditional density estimators, without imposing specific convergence rates. For example, the propensity score estimator of Shao and Wang (2016) satisfies Assumption 2 under standard conditions for kernel estimation. Also, the conditional density estimator in (2.8) satisfies Assumption 3 under standard regularity conditions, as supported by existing results for linear quantile regression

(Wang et al., 2019) and quantile random forests (Meinshausen, 2006). Assumption 4 ensures that the conditional CDF of the nonconformity score is continuous with controlled local variation. This helps reduce the discrepancy between the CDF and its estimate. Assumption 5 ensures that the local region is small to accurately approximate the conditional distribution, while containing enough data points to support reliable estimation.

Theorem 3. *Under Assumptions 2–5, the localized prediction in Algorithm 2 converges to the HPD set,*

$$\mathbb{P}\{Y_{n+1} \in C(\mathbf{X}_{n+1};\alpha)\Delta\widehat{C}(\mathbf{X}_{n+1};\alpha)\} = o(1),$$

where Δ *denotes symmetric set difference, i.e.,* $A\Delta B := (A\cap B^c)\cup(B\cap A^c)$. *Furthermore,* $\widehat{C}(\mathbf{X}_{n+1};\alpha)$ *satisfies the asymptotic conditional coverage, that is, there exists a sequence of (possibly random) sets* $\Lambda_n \subset \mathcal{X}$ *such that* $\mathbb{P}(\mathbf{X}_{n+1} \in \Lambda_n) = 1 - o(1)$ *and*

$$\sup_{\mathbf{x}_{n+1}\in\Lambda_n} \left|\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(\mathbf{X}_{n+1};\alpha) \mid \mathbf{X}_{n+1} = \mathbf{x}_{n+1}\right\} - (1-\alpha)\right| = o(1).$$

Theorem 3 shows that the proposed prediction set converges to the oracle prediction set, implying that it is asymptotically optimal and achieves the smallest Lebesgue measure. Furthermore, it guarantees asymptotic conditional coverage for any given $\mathbf{x}_{n+1}$, a much stronger result than Theorems 1 and 2, which establish only average coverage over a region of the covariate

Statistica Sinica: Newly accepted Paper

space. This highlights the improved efficiency of our method. The proof is referred to the supplementary materials.

4. Simulation Study

This section evaluates the finite-sample performance of the proposed methods in terms of marginal coverage, with conditional coverage further examined in Section S2 of the supplementary materials.

For comparison, we evaluate the performance of the following methods: (1) **OMNI**: the omniscient method, which applies standard conformal prediction using the true responses Y for all observations, with no missingness. This method serves as a gold standard but is unachievable in real-world settings. (2) **Naive**: standard conformal prediction applied to fully observed data, i.e., excluding data points with missing responses. (3) **MAR-CP**: weighted conformal prediction based on MAR weights. This is implemented by setting $q\left(Y, \boldsymbol{\gamma}\right) = 1$ in the propensity (2.12). (4) **MNAR-CP**: the proposed weighted conformal prediction using MNAR-based weights.

We consider two models: a linear model and a nonlinear model. In both cases, the covariate vector $\mathbf{X} = (\mathbf{Z}, \mathbf{U})$ includes a discrete instrumental variable $\mathbf{Z}$, where $\Pr(Z_i = 1) = 0.2$, $\Pr(Z_i = 2) = 0.4$, and $\Pr(Z_i = 3) = 0.4$. The remaining covariates, $U_{ij} \sim N(Z_i/10, 1)$ for $1 \leq j \leq 10$. The missingness indicator follows $\delta \sim \text{Bernoulli}(\pi)$. The outcome and missingness are generated as follows. (1) In the **linear model**, $Y_i = Z_i + \sum_{j=1}^{10} U_{ij} + \epsilon_i$ and $\pi_i = \{1 + \exp(a - 0.1\sum_{j=1}^{10} U_{ij} + 0.65\,Y_i)\}^{-1}$. (2) In the **nonlinear model**, $Y_i = Z_i + 0.5\big(\sum_{j=5}^{10} U_{ij}\big)^2 + \epsilon_i$ and $\pi_i = \{1+\exp(a-0.1(\sum_{j=5}^{10} U_{ij})^2+0.4\,Y_i)\}^{-1}$. We consider both **homoscedastic** errors $\epsilon_i \sim N(0,1)$ and **heteroscedastic** errors $\epsilon_i \sim \text{Gamma}(0.5Z_i, 0.5Z_i)$. The parameter a is set as a =-3.5, -3, -2.5, which correspond to missing rates of approximately 40%, 50%, and 60% in the linear model, and 20%, 30%, and 40% in the nonlinear model.

We focus on the non-localized prediction from Algorithm 1, which is sufficient for evaluating marginal coverage. The localized prediction from Algorithm 2 is evaluated separately in Section S2 of the supplementary materials. In our non-localized prediction, we set $\mathcal{Y}_{\text{grid}}$ to be a grid of 200 uniformly spaced points over the empirical range of the observed responses. The weight is obtained from equation (2.11), where the propensity score $\mathbb{P}(\delta = 1 \mid \mathbf{x}, y)$ is estimated using the method of Wang et al. (2021). The conditional density is estimated using equation (2.8), where $f(y \mid \mathbf{x}, \delta = 1)$ is obtained via the quotient method (2.10) with quantile levels $\{1/32, \ldots, 31/32\}$. The bandwidth h_n is chosen using the *bandwidth.rq* function from the R package *quantreg*, ensuring that h_n converges to zero at the rate of $n^{-1/3}$. To estimate the conditional quantile $Q_Y(\tau \mid \mathbf{x}, \delta = 1)$,

Statistica Sinica: Newly accepted Paper

we employ two methods: linear quantile regression (**LQR**) and quantile random forest (**QRF**), implemented using the R packages *quantreg* and *grf*, respectively. For the adjustment factor $\rho(\mathbf{x}, y)$ in (2.8), we estimate $\mathbb{P}(\delta = 1 \mid \mathbf{x})$ using the gradient boosting method via the R package *gbm*, and directly use the estimate of $\mathbb{P}(\delta = 1 \mid \mathbf{x}, y)$ from the weight estimation.

We repeat the simulation 500 times. In each replication, we generate datasets with $n = 1000$ and $n = 4000$ samples to construct prediction sets with 90% nominal coverage, while evaluate their performance on an additional 500 data points $(\mathbf{X}_{n+1}, Y_{n+1}), \ldots, (\mathbf{X}_{n+500}, Y_{n+500})$. The coverage probabilities and interval lengths for different models and sample sizes are shown in Tables 1–2, with part of the results deferred to the supplementary material for brevity. It can be seen that the proposed MNAR-CP method nearly maintains the nominal 90% level, with standard errors comparable to those of OMNI when $n = 4000$. However, the Naive and MAR-CP methods consistently exhibit undercoverage, as they do not correct for the joint effect of missingness on both $\mathbf{X}$ and Y. As a result, even when MAR-CP produces intervals with lengths comparable to those of MNAR-CP, its biased weights lead to miscalibrated conformity scores, causing the method to fail to achieve nominal coverage.

Since Naive and MAR-CP do not guarantee valid coverage, we do not

Statistica Sinica

Statistica Sinica: Newly accepted Paper

Table 1: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets, across different missing rates and $\widehat{f}$ estimators, averaged over 500 new subjects and 500 repetitions in linear model with $n = 1000$.

	Miss.(%)	$\widehat{f}$	OMNI	Naive	MAR-CP	MNAR-CP
				(a) Homoskedastic		
AC% (SE×100)	40	LQR	89.85(0.08)	88.36(0.11)	87.56(0.29)	**89.41(0.12)**
		QRF	89.95(0.08)	79.84(0.14)	81.98(0.25)	**88.60(0.19)**
	50	LQR	89.85(0.08)	87.98(0.13)	86.54(0.46)	**89.39(0.12)**
		QRF	90.02(0.08)	77.77(0.15)	80.53(0.26)	**89.09(0.19)**
	60	LQR	89.85(0.08)	87.59(0.15)	85.78(0.51)	**89.46(0.15)**
		QRF	89.98(0.08)	75.28(0.18)	79.17(0.32)	**88.19(0.22)**
AL (SE)	40	LQR	3.46(0.01)	3.51(0.01)	3.71(0.07)	4.36(0.09)
		QRF	8.16(0.01)	7.56(0.02)	8.23(0.09)	8.71(0.07)
	50	LQR	3.46(0.01)	3.55(0.01)	3.83(0.08)	4.90(0.11)
		QRF	8.16(0.01)	7.56(0.02)	8.27(0.08)	8.75(0.07)
	60	LQR	3.46(0.01)	3.62(0.01)	3.89(0.05)	5.64(0.14)
		QRF	8.16(0.01)	7.57(0.02)	8.51(0.10)	8.87(0.08)
				(b) Heteroscedastic		
AC% (SE×100)	40	LQR	90.07(0.09)	83.97(0.14)	83.48(0.34)	**89.17(0.17)**
		QRF	90.12(0.09)	76.82(0.16)	80.17(0.29)	**88.79(0.20)**
	50	LQR	90.07(0.09)	83.12(0.16)	81.93(0.47)	**88.78(0.19)**
		QRF	90.02(0.09)	74.18(0.18)	78.61(0.35)	**88.13(0.22)**
	60	LQR	90.07(0.09)	82.68(0.19)	81.40(0.50)	**89.10(0.21)**
		QRF	89.99(0.09)	71.70(0.22)	77.84(0.41)	**88.27(0.24)**
AL (SE)	40	LQR	2.76(0.01)	2.31(0.01)	2.74(0.09)	6.39(0.22)
		QRF	8.19(0.02)	7.45(0.02)	8.35(0.10)	8.80(0.08)
	50	LQR	2.76(0.01)	2.29(0.01)	2.95(0.11)	7.27(0.23)
		QRF	8.17(0.02)	7.45(0.02)	8.58(0.11)	9.05(0.09)
	60	LQR	2.76(0.01)	2.45(0.05)	3.30(0.13)	8.67(0.24)
		QRF	8.17(0.02)	7.53(0.02)	9.02(0.13)	9.14(0.09)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

Table 2: Average Coverage percentages (AC%) and Average Length (AL) for 90% prediction sets, across different missing rates and $\widehat{f}$ estimators, averaged over 500 new subjects and 500 repetitions in nonlinear model with $n = 4000$.

	Miss.(%)	$\widehat{f}$	OMNI	Naive	MAR-CP	MNAR-CP
	(a) Homoskedastic					
AC% (SE×100)	20	LQR	89.97(0.06)	84.11(0.08)	86.48(0.23)	**90.00(0.10)**
		QRF	89.93(0.07)	84.22(0.08)	86.64(0.23)	**89.81(0.09)**
	30	LQR	89.97(0.06)	82.88(0.08)	85.65(0.24)	**90.00(0.11)**
		QRF	89.94(0.07)	82.91(0.09)	85.77(0.25)	**89.82(0.10)**
	40	LQR	89.97(0.06)	81.40(0.09)	84.80(0.28)	**89.67(0.11)**
		QRF	89.96(0.07)	81.32(0.09)	84.82(0.28)	**89.52(0.10)**
AL (SE)	20	LQR	10.71(0.02)	8.55(0.02)	12.08(0.43)	12.56(0.27)
		QRF	10.93(0.01)	9.27(0.01)	11.82(0.34)	10.94(0.03)
	30	LQR	10.71(0.02)	8.20(0.02)	11.95(0.45)	13.96(0.35)
		QRF	10.92(0.02)	8.96(0.01)	11.82(0.37)	10.93(0.03)
	40	LQR	10.71(0.02)	7.86(0.02)	12.22(0.48)	15.31(0.44)
		QRF	10.93(0.02)	8.64(0.01)	11.79(0.39)	10.90(0.03)
	(b) Heteroscedastic					
AC% (SE×100)	20	LQR	90.08(0.07)	82.57(0.09)	85.06(0.24)	**89.83(0.11)**
		QRF	90.07(0.07)	82.55(0.09)	85.10(0.25)	**89.61(0.10)**
	30	LQR	90.08(0.07)	80.96(0.09)	83.89(0.27)	**89.51(0.11)**
		QRF	90.06(0.07)	80.80(0.09)	83.91(0.27)	**89.24(0.10)**
	40	LQR	90.08(0.07)	79.18(0.09)	82.47(0.38)	**89.18(0.12)**
		QRF	90.04(0.07)	78.87(0.10)	82.53(0.32)	**88.67(0.11)**
AL (SE)	20	LQR	10.13(0.02)	7.50(0.01)	10.86(0.42)	13.02(0.37)
		QRF	10.48(0.01)	8.43(0.01)	10.91(0.33)	10.33(0.03)
	30	LQR	10.13(0.02)	7.13(0.01)	10.56(0.42)	14.54(0.46)
		QRF	10.47(0.01)	8.08(0.01)	10.49(0.32)	10.26(0.03)
	40	LQR	10.13(0.02)	6.73(0.02)	11.02(0.50)	15.51(0.49)
		QRF	10.47(0.01)	7.70(0.01)	10.93(0.40)	10.09(0.03)

OMNI: Standard conformal prediction applied to the complete data. Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

further discuss their interval lengths. Our proposed method, MNAR-CP, yields slightly wider intervals than OMNI. This is expected because OMNI uses fully observed outcomes with equal weights, whereas MNAR-CP relies on the estimated weights, which increase variability and reduce the effective sample size, resulting in wider intervals. The variation across the estimators $\widehat{f}$ reflects the model-free property of our method with respect to the regression model $Y \mid \mathbf{X}$. The LQR-based $\widehat{f}$ is correctly specified under the linear setting (Table 1) but misspecified under the nonlinear setting (Table 2), whereas the QRF-based $\widehat{f}$ is a flexible nonparametric estimator that does not rely on parametric model assumptions. In both cases, our method achieves valid coverage, with the misspecified method yielding longer prediction intervals in order to maintain coverage. In addition, Figure S2 in the Supplementary Material provides a visual comparison of density estimates across different regression model specifications.

5. Analysis of the ACTG 175 Data

In this section, we apply our proposed method to a dataset from the AIDS Clinical Trials Group Protocol 175 (ACTG 175; Hammer et al., 1996), which is available in the R package *speff2trial*. Specifically, this dataset includes 2,139 HIV-infected patients who were randomly divided into four

Table 3: Description of covariates in the ACTG175 data.

Variable	**Description**
gender	Male and female
race	White versus non-white
wtkg	Weight in kilogram
days	Days until the first occurrence of: (i) a CD4 count decline of at least 50, (ii) an AIDS progression event, or (iii) death
cd40	CD4 count (cells/mm^3) at baseline
cd420	CD4 count (cells/mm^3) around 20 weeks after treatments
cd80	CD8 count (cells/mm^3) at baseline
cd820	CD8 count (cells/mm^3) around 20 weeks after treatments

groups based on the regimen received: (I) zidovudine (ZDV) monotherapy with 532 subjects; (II) ZDV + didanosine (ddI) with 522 subjects; (III) ZDV + zalcitabine with 524 subjects; and (IV) ddI monotherapy with 561 subjects. For illustrative purposes, we consider only the patients in group (I); the analyses for the other groups are similar.

To evaluate the effectiveness of HIV treatment, a key strategy is to monitor CD4 cell counts in HIV-positive patients, with increases typically indicating improved health. Consequently, a practical problem is to predict the CD4 cell count of a new patient following treatment. Let Y denote the CD4 cell count measured approximately 96 weeks after treatment, with 39.66% of the observations missing due to loss to follow-up. We assume that this missingness is related to its underlying value and is therefore nonignorable, as Hogan and Laird (1997) and Yuan and Yin (2010) found

Statistica Sinica

Table 4: Average Coverage percentage (AC%) and Average Length (AL) for 90% prediction sets, across different $\widehat{f}$ estimators, averaged over leave-one-out cross-validation on 321 observed subjects.

	$\widehat{f}$	Naive	MAR-CP	MNAR-CP
AC%(SE×100)	LQR	92.52(0.39)	91.90(0.42)	90.03(0.50)
	QRF	90.03(0.50)	90.65(0.47)	91.28(0.44)
AL(SE)	LQR	366.35(8.44)	382.76(10.55)	**352.17(9.03)**
	QRF	352.44(3.89)	363.53(7.03)	**308.93(7.14)**

Naive: Standard conformal prediction applied to the observed data. MAR-CP: Weighted conformal prediction with MAR weights. MNAR-CP: Proposed weighted conformal prediction with MNAR weights. Values in parentheses are standard errors.

that patients with low CD4 cell counts are more likely to miss scheduled study visits than those with normal counts. To determine the covariates $\mathbf{X}$, we consider the variables used in Zhao et al. (2021) and Li et al. (2023), while further incorporating the importance measure function from Athey et al. (2019) to assess the importance of variables with respect to Y. Finally, we selected 8 covariates as detailed in Table 3. We adopt a semiparametric propensity score model and, following Wang et al. (2021), use gender and race as instrumental variables $\mathbf{Z}$, assuming that the missingness propensity is conditionally independent of $\mathbf{Z}$ given Y and other covariates $\mathbf{U}$.

We compare three methods, Naive, MAR-CP, and MNAR-CP, under the same weight and density estimation settings as in Section 4. Method

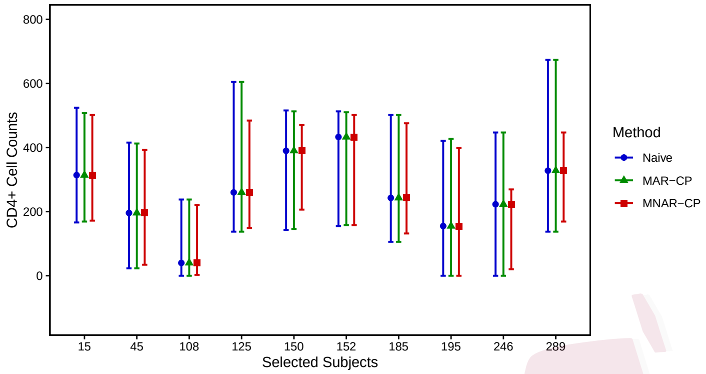

Figure 1: Comparison of 90% prediction intervals by Naive, MAR-CP and MNAR-CP methods for 10 randomly selected subjects. The points show the true values, and the horizontal axis shows the subject indices.

performance is evaluated via leave-one-out cross-validation, where each subject is treated as the test case and the remaining samples are used for training and calibration. To calculate coverage, we use only the 321 subjects with observed responses, and the results are reported in Table 4. MNAR-CP achieves nominal coverage while producing significantly shorter intervals, particularly with the QRF estimator, thereby demonstrating greater efficiency. Although the Naive and MAR-CP methods also attain nominal coverage, this reflects accuracy only on the observed distribution, because coverage is computed over observed cases and unobserved cases cannot be evaluated without ground truth.

To more comprehensively evaluate predictive performance, we compare

Statistica Sinica: Newly accepted Paper

the interval lengths for each subject individually across the three methods, focusing on QRF for conditional density estimation because it is more effective than LQR. We found that 87.82% of the intervals from MNAR-CP are shorter than those from Naive, and 89.70% are shorter than those from MAR-CP. Additionally, we present the prediction intervals for 10 randomly selected patients in Figure 1, further illustrating that MNAR-CP provides more efficient predictions for the majority.

6. Discussion

In this paper, we go beyond classical statistical inference for parameters under nonignorable missing data to develop a new approach for uncertainty quantification. We propose both non-localized and localized MNAR-weighted conformal prediction frameworks for constructing prediction sets for new test data, where the localized version is applied when personalized inference and conditional coverage at each point are desired. Importantly, we also introduce a novel conditional density estimation that leverages the observed data to identify the density function along the quantile process. This conditional density estimator can be readily incorporated into our framework, enabling the resulting predictions to achieve not only approximate marginal coverage but also local and asymptotic conditional coverage.

There are other interesting directions for future research. Our method is robust to misspecification of the conditional density, but it requires the propensity score to be consistently estimated. This model reliance may be relaxed by adopting alternative identification strategies that impose additional structural assumptions. For example, Li et al. (2023) propose a nonparametric estimation framework based on a representer equation and a shadow variable assumption. In addition, it is meaningful to conduct sensitivity analyses for conformal prediction under missing data mechanisms that violate the MAR assumption. One possible approach is to introduce an odds ratio–based sensitivity model (Jin et al., 2023; Yin et al., 2024).

Supplementary Materials

The online Supplementary Materials provided detailed technical proofs, high-dimensional extensions, and additional numerical experiments.

Acknowledgements

We sincerely thank the Editor, Associate Editor, and two anonymous reviewers for their constructive comments and helpful suggestions. The research of Tang and Zhang was partially supported by the National Natural Science Foundation of China (grants 12371265 and 12471280), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry

Statistica Sinica

Statistica Sinica: Newly accepted Paper

of Education of China (grant JYB2025XDXM904) and Chern Institute of Mathematics, Nankai University, and the research of Wang was partially supported by the National Science Foundation (grant DMS-2436216).

References

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

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

Candès, E., L. Lei, and Z. Ren (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(1), 24–45.

Chen, W., K.-J. Chun, and R. F. Barber (2018). Discretized conformal prediction for efficient distribution-free inference. *Stat 7*(1), e173.

Gui, Y., R. Hore, Z. Ren, and R. F. Barber (2024). Conformalized survival analysis with adaptive cut-offs. *Biometrika 111*(2), 459–477.

Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine 335*(15), 1081–1090.

Hogan, J. W. and N. M. Laird (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine 16*(3), 259–272.

Izbicki, R., G. Shimizu, and R. B. Stern (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *The Journal of Machine Learning Research 23*(1), 3772–3803.

Jin, Y., Z. Ren, and E. J. Candès (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences 120*(6), e2214889120.

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

Statistica Sinica

Statistica Sinica: Newly accepted Paper

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika 106*(4), 749–764.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*(523), 1094–1111.

Lei, J. and L. Wasserman (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(1), 71–96.

Lei, L. and E. J. Candès (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(5), 911–938.

Li, M., Y. Ma, and J. Zhao (2022). Efficient estimation in a partially specified nonignorable propensity score model. *Computational Statistics & Data Analysis 174*, 107322.

Li, P., J. Qin, and Y. Liu (2023). Instability of inverse probability weighting methods and a remedy for nonignorable missing data. *Biometrics 79*(4), 3215–3226.

Li, W., W. Miao, and E. Tchetgen Tchetgen (2023). Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(3), 913–935.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis With Missing Data*, Volume 793. John Wiley & Sons.

Liu, Y., P. Li, and J. Qin (2022). Full-semiparametric-likelihood-based inference for non-ignorable missing data. *Statistica Sinica 32*(1), 271–292.

Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research 7*(35), 983–999.

Miao, W., P. Ding, and Z. Geng (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association 111*(516), 1673–1683.

Miao, W., L. Liu, Y. Li, E. J. Tchetgen Tchetgen, and Z. Geng (2024). Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM IMS Journal of Data Science 1*(2), 1–23.

Qiu, Z., C. Peng, Y. Tang, and H. J. Wang (2026). Review of recent advances in high-dimensional quantile regression. *Wiley Interdisciplinary Reviews: Computational Statistics*. To appear.

Statistica Sinica: Newly accepted Paper

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine 16*(3), 285–319.

Shafer, G. and V. Vovk (2008). A tutorial on conformal prediction. *The Journal of Machine Learning Research 9*(12), 371–421.

Shao, J. and L. Wang (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika 103*(1), 175–187.

Siddiqui, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *Journal of Research of the National Bureau of Standards-B 64*, 145–150.

Sun, B., W. Miao, and D. S. Wickramarachchi (2026). On doubly robust estimation with nonignorable missing data using instrumental variables. *Statistica Sinica 36*(4). To appear.

Tan, K. M., L. Wang, and W.-X. Zhou (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 205–233.

Tian, Q., D. Zeng, and J. Zhao (2025). Identification and efficient estimation in regression analysis with response missing not at random. *Statistica Sinica*. Forthcoming.

Tibshirani, R. J., R. Foygel Barber, E. Candes, and A. Ramdas (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems 32*.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*, Volume 29. Springer.

Wang, H. J., X. Feng, and C. Dong (2019). Copula-based quantile regression for longitudinal data. *Statistica Sinica 29*(1), 245–264.

Wang, L., P. Zhao, and J. Shao (2021). Dimension-reduced semiparametric estimation of distribution functions and quantiles with nonignorable nonresponse. *Computational Statistics & Data Analysis 156*, 107142.

Yin, M., C. Shi, Y. Wang, and D. M. Blei (2024). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association 119*(545), 122–135.

Yu, A., Y. Zhong, X. Feng, and Y. Wei (2023). Quantile regression for nonignorable missing data with its application of analyzing electronic medical records. *Biometrics 79*(3), 2036–2049.

Statistica Sinica: Newly accepted Paper

Yuan, Y. and G. Yin (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics 66*(1), 105–114.

Zhang, L., C. Lin, and Y. Zhou (2018). Generalized method of moments for nonignorable missing data. *Statistica Sinica 28*(4), 2107–2124.

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

Zhang, Y., C. Shi, and S. Luo (2023). Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 2751–2768. PMLR.

Zhao, J. and Y. Ma (2022). A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association 117*(540), 1916–1930.

Zhao, J. and J. Shao (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association 110*(512), 1577–1590.

Zhao, P., N. Tang, and H. Zhu (2020). Generalized empirical likelihood inferences for nonsmooth moment functions with nonignorable missing values. *Statistica Sinica 30*(1), 217–249.

Zhao, P., L. Wang, and J. Shao (2021). Sufficient dimension reduction and instrument search for data with nonignorable nonresponse. *Bernoulli 27*(2), 930–945.

KLATASDS-MOE, School of Statistics, East China Normal University

E-mail: menghany@umich.edu

Academy of Statistics and Interdisciplinary Sciences, East China Normal University

E-mail: yyzhang@fem.ecnu.edu.cn

KLATASDS-MOE, School of Statistics, East China Normal University

E-mail: yltang@fem.ecnu.edu.cn * Corresponding author

Department of Statistics, Rice University

E-mail: jw322@rice.edu * Corresponding author