Statistica Sinica Preprint No: SS-2025-0389	
Title	Conformal Inference for Missing Data Under Multiple Robust Learning
Manuscript ID	SS-2025-0389
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0389
Complete List of Authors	Wenlu Tang, Hongni Wang, Xingcai Zhou, Bei Jiang and Linglong Kong
Corresponding Authors	Linglong Kong
E-mails	lkong@ualberta.ca
Notice: Accepted author version.	

Statistica Sinica

CONFORMAL INFERENCE FOR MISSING DATA UNDER MULTIPLE ROBUST LEARNING

Wenlu Tang[1*], Hongni Wang[2*], Xingcai Zhou[3], Bei Jiang[1] and Linglong Kong[1]

[1]*University of Alberta,* [2]*Shandong University of Finance and Economics, and* [3]*Nanjing Audit University*

Abstract: We develop a novel approach to address the common but challenging problem of conformal inference for missing data in machine learning, focusing on Missing at Random (MAR) data. We propose a new procedure, *Conformal Prediction for Missing Data under Multiple Robust Learning (CM–MRL)*, which combines split conformal calibration with a multiple robust empirical-likelihood (EL) reweighting scheme. The method proceeds via a double calibration by reweighting the complete-case scores by EL so that their distribution matches the full calibration distribution implied by MAR, even when some working models are misspecified. We demonstrate the asymptotic behavior of our estimators through empirical process theory, provide reliable coverage for our prediction intervals, both marginally and conditionally, and further show an interval-length dominance result. We demonstrate the effectiveness of the proposed method through several numerical experiments in the presence of missing data.

[*]These authors contributed equally to this work.

Key words and phrases: Conformal inference, Missing data, Multiple robust model, Uncertainty quantification, Missing at random.

1. Introduction

In many statistical applications, quantifying predictive uncertainty is crucial. Prediction intervals indicate a range in which a future outcome is expected to lie with a specified probability; for example, for data $(X_i, Y_i)_{i=1}^n$, the goal is to create a prediction set $\hat{C}(X_{n+1})$ that includes Y_{n+1} with probability at least $1-\alpha$, where α is the miscoverage rate. However, the integrity of these predictive intervals is frequently challenged by a common but significant issue in data analysis: the presence of missing values (Little and Rubin, 2002), which is pervasive, spanning various disciplines.

In statistical analysis, missing data introduces biases and uncertainties that can lead to skewed inferences. Broadly classified into Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), each type has unique challenges and tailored strategies. In the missing at random (MAR) setting, for pairwise data $(X_i, Y_i)_{i=1}^n$, the response Y is observed when $R = 1$ and unobserved, i.e., missing, when $R = 0$, and $\mathbb{P}(R = 1 \mid Y, X) = \mathbb{P}(R = 1 \mid X)$. This missing-data mechanism induces selection bias, and naive prediction intervals that

ignore missingness may be invalid. Two classical approaches to handle missingness are inverse probability weighting (IPW) and imputation. IPW reweights each observed outcome by the inverse of an estimated propensity score $\pi(X) = P(R = 1 \mid X)$ (Rosenbaum and Rubin, 1983; Tu et al., 2019), whereas imputation predicts missing Y from observed covariates, e.g. using regression or multiple imputation (Rubin, 1996, 2018). Both methods can yield consistent point estimates of quantities like $\mathbb{E}[Y \mid X]$ when models are correctly specified. However, both methods are sensitive to model misspecification, which can result in biased estimators (Little and Rubin, 2019). To address this, prior studies have suggested a doubly robust approach (Robins and Rotnitzky, 1995; Cao et al., 2009; Tan, 2010), ensuring consistency if either the propensity-score model or the outcome model is correct. However, such approaches can still fail if both models are wrong, since correctly specifying these models is challenging. Prediction intervals based on potentially fragile imputations require additional robustness. Consequently, the concept of multiple robust learning has emerged for handling missing data.

1.1 Conformal Inference

Interval estimation can quantify the uncertainty of a point estimate. Methods based on conformal prediction have been developed for distribution-free

interval estimation in recent years. Given pairwise observations $(X_i, Y_i)_{i=1}^n$ and miscoverage rate α, the goal of conformal prediction is to construct an interval $C(x)$ such that, for a new observation (X_{n+1}, Y_{n+1}), $\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha$. Since the pioneering work by Vovk et al. (2005), there have been numerous conformal prediction studies and extensions in both computation and theory, see Romano et al. (2019), Tibshirani et al. (2019) and Zaffran et al. (2022). The conformal inference method can be applied to various scenarios, including time series (Zaffran et al., 2022), survival data (Candès et al., 2023), causal inference (Lei and Candès, 2021) and distribution shifts (Gibbs and Candes, 2021). Marginal validity, a conventional coverage guarantee that can be achieved under the i.i.d. assumption, is demonstrated in Lei et al. (2013) and Lei and Wasserman (2014). However, as demonstrated in Lei et al. (2018) and Vovk (2012), conditional validity with a finite-length prediction interval is impossible without regularity and consistency assumptions on the model and estimator. The conditional validity is hard to achieve in missing data prediction intervals since methods for missing data rely on correctly specified models.

Zaffran et al. (2023) introduced a framework for conformalized quantile regression, showing that missing-data augmentation with the *impute-then-predict* strategy is marginally valid, and explored conditional validity within

a masked framework. However, their research primarily addresses missing covariate values, and the accuracy of the conformal interval is highly influenced by the choice of imputation model. Therefore, we propose a conformal interval that is adaptive to missing responses and less sensitive to imputation models.

1.2 Multiple Robust Estimation

Suppose the observations are pairwise samples $\{R_iY_i, X_i, R_i\}_{i=1}^n$, where Y is the response variable, $X \in \mathbb{R}^p$ is the vector of covariates, and R_i is the indicator of missingness. We write $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing. In this paper, we consider the typical missing at random (MAR) mechanism (Little and Rubin, 2002); that is,

$$\mathbb{P}(R = 1 \mid Y, X) = \mathbb{P}(R = 1 \mid X),$$

and this conditional probability is called the propensity score function, denoted by $\pi(X)$. Let $m = \sum_{i=1}^n R_i$ be the number of fully observed samples. Without loss of generality, assume that the indices of the completely observed samples with $R_i = 1$ are $i = 1, \ldots, m$. For regression tasks with missing responses, the parameter of interest is $\mu_0(x) = \mathbb{E}(Y|X = x)$ for mean regression or $\mu_\tau(x) = Q_\tau(Y \mid X = x)$ for quantile regression. Multiple robust learning introduces a general framework for deriving a multiple

robust estimator of $\mu(x)$.

Previous studies on multiple robust estimation for mean Han and Wang (2013) and quantile Han et al. (2019) regression problems can be summarized in a general framework that incorporates multiple propensity and error-imputation models.

1.3 Related Works

Multiple robust (MR) approaches reduce sensitivity to model misspecification by combining multiple propensity-score and imputation models, yielding consistent estimation if any one working model is correct (Han and Wang, 2013). These multiple robust approaches involve both imputation and inverse probability weighting (IPW) (Han, 2014b). By leveraging information from multiple models, the calibrated empirical-likelihood estimator attains the semiparametric efficiency bound when both the propensity-score model and an outcome model are correctly specified (Han, 2016). Such multiple robust methods can mitigate sensitivity to near-zero estimates in IPW methods (Kang and Schafer, 2007) and reduce bias in imputation approaches. The multiple robust approach can be applied to several missing-data and regression scenarios, including marginal quantile/mean estimation (Han et al., 2019) and regression with missing responses/covariates (Li

et al., 2020). The computational cost lies in the constrained optimization, which can be solved via the Newton-Raphson method (Han, 2014a).

The literature on multiple robust estimation develops a general robust estimation framework for missing data, using calibrated empirical likelihood weights to achieve consistency and efficiency when at least one working propensity or outcome model is correct. In contrast, our goal is to construct valid prediction intervals under the MAR missingness mechanism, which requires further calibrating the distribution of conformity scores rather than only estimating conditional means or quantiles.

1.4 Summary

In this paper, we develop a robust conformal prediction method for MAR missing data with both methodological and theoretical contributions:

- We propose a multiple-robust regression framework that combines multiple propensity-score and outcome models, yielding consistency under MAR if any one working model is correct.

- We introduce a double calibration procedure for conformal intervals: reweight complete-case conformity scores using multiple-robust empirical likelihood weights, and calibrate score quantiles by pooling across multiple imputations to account for missingness and imputa-

tion uncertainty.

- We establish asymptotic marginal and conditional validity under standard conditions, and demonstrate strong finite-sample performance in simulations and a real-data application, often producing shorter intervals than naive alternatives.

Section 2.1 presents the MR estimator, Section 2.2 develops the double calibration framework, and Algorithm 1 summarizes the *Conformal Prediction for Missing Data under Multiple Robust Learning (CM–MRL)* procedure. Sections 3 and 4 provide asymptotic results and empirical evaluations.

2. Methodology

2.1 Multiple Robust Estimation

In this section, we propose the multiple robust estimator for the CM-MRL interval construction following the main procedure of the MRL in Han (2014b); Han et al. (2019). Let the full sample set be $D_i = (X_i, R_i, R_iY_i)$, $i = 1, \ldots, n$, where $R_i \in \{1, 0\}$ is the response-observation indicator. In practice, we fully observe only $\sum_{i=1}^{n} R_i = m$ pairs. In the missing at random (MAR) setting, the response Y is observed when $R = 1$ and missing when $R = 0$, and $\mathbb{P}(R = 1 \mid Y, X) = \mathbb{P}(R = 1 \mid X)$. We adopt a multiple ro-

bust method to derive a consistent estimator for this regression task with missing responses. Specifically, one can consider J propensity models with parameter a and K imputation models with parameter b:

$$\Pi = \{\pi^1(\hat{a}^1; X), \ldots, \pi^J(\hat{a}^J; X)\}; \quad \mathcal{F} = \{f^1(Y \mid X; \hat{b}^1), \ldots, f^K(Y \mid X; \hat{b}^K)\}.$$

Initial estimators of the parameters $\{\hat{a}^1, \ldots, \hat{a}^J\}$ and $\{\hat{b}^1, \ldots, \hat{b}^K\}$ in the working models are obtained by maximum likelihood. We maximize the binomial likelihood to obtain the estimator $\hat{a}^j$ for the propensity score,

$$\prod_{i=1}^{n} \left\{\pi^j\left(a^j; X_i\right)\right\}^{R_i} \left\{1 - \pi^j\left(a^j; X_i\right)\right\}^{1-R_i}. \tag{2.1}$$

Next we obtain the estimator $\hat{b}^k$ by maximizing the conditional likelihood:

$$\prod_{i=1}^{n} \left\{f^k\left(Y_i \mid \boldsymbol{X}_i; b^k\right)\right\}^{R_i}. \tag{2.2}$$

Next, we can impute the missing values by random sampling from the conditional distribution $f^k\left(Y_i \mid \boldsymbol{X}_i; \hat{b}^k\right)$. Let $L(Y-\mu(x))$ be a loss function, and let $Y_i^t(\hat{b}^k)$ be the t-th random draw from the k-th imputation model. By averaging the T repeated random draws, the imputed values can be more robust to extreme random draws. Subsequently, we obtain the estimator $\hat{\mu}_k(x)$ under each imputation model k by minimizing

$$\ell(\mu_k) = \sum_{i=1}^{n} R_i L\left(Y_i - \mu_k(X_i)\right) + \sum_{i=1}^{n} \left(1 - R_i\right) \frac{1}{T} \sum_{t=1}^{T} L\left\{Y_i^t\left(\hat{b}^k\right) - \mu_k(X_i)\right\}. \tag{2.3}$$

The estimator $\hat{\mu}_k(x)$ from every imputed dataset usually introduces bias, as the imputation model may not perfectly represent the true model $f(Y \mid X = x)$. To correct the bias, we can apply weights, denoted by w, to balance the fully observed samples and the samples with imputed data.

Let $g(\mu) = \partial L(\mu)/\partial\mu$ be the first derivative of the loss function with respect to μ, and define $g^k(\hat{b}^k; X_i) = T^{-1}\sum_{t=1}^{T} g\{Y_i^t\left(\hat{b}^k\right) - \hat{\mu}_k(x)\}$. The weight w_i plays a role similar to inverse-probability weights (IPW) for fully observed samples. In robust estimation involving multiple models, the weights w are applied in a way that satisfies

$$\sum_{i=1}^{m} w_i \pi^j\left(\hat{a}^j; X_i\right) = \hat{\theta}^j = n^{-1}\sum_{i=1}^{n} \pi^j\left(\hat{a}^j; X_i\right), \tag{2.4}$$

$$\sum_{i=1}^{m} w_i g^k\left(\hat{b}^k; X_i\right) = \hat{\eta}^k = n^{-1}\sum_{i=1}^{n} g^k\left(\hat{b}^k; X_i\right), \tag{2.5}$$

with the constraint $\sum_{i=1}^{m} w_i = 1$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Let $\hat{\boldsymbol{a}} = (\hat{a}^1, \ldots, \hat{a}^J)$, $\hat{\boldsymbol{b}} = (\hat{b}^1, \ldots, \hat{b}^K)$, $\hat{\boldsymbol{\mu}} = (\hat{\mu}^1, \ldots, \hat{\mu}^K)$, and let $\hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\mu}})$ be the normalized vector for multiple models of dimension $J + K$:

$$\begin{aligned}\hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\mu}}) = \Big\{ & \pi^1\left(\hat{a}^1; X_i\right) - \hat{\theta}^1, \ldots, \pi^J\left(\hat{a}^J; X_i\right) - \hat{\theta}^J, \\ & g^1\left(\hat{b}^1; X_i\right) - \hat{\eta}^1, \ldots, g^K\left(\hat{b}^K; X_i\right) - \hat{\eta}^K \Big\}^T.\end{aligned}$$

For every fully observed sample $i = 1, \ldots, m$, the empirical-likelihood (EL) weights $\{\hat{w}_i\}_{i:R_i=1}$ are then obtained by

$$\hat{w}_i = \frac{1}{m} \frac{1}{1 + \hat{\boldsymbol{\rho}}^T \hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\mu}})}, \tag{2.6}$$

where $\hat{\rho}$ minimizes

$$F_m(\boldsymbol{\rho}) = -\frac{1}{m} \sum_{i=1}^{m} \log \left\{ 1 + \boldsymbol{\rho}^T \hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\mu}}) \right\}. \tag{2.7}$$

The nonnegativity of $\hat{w}_i$ imposes the feasibility constraints

$$1 + \hat{\boldsymbol{\rho}}^T \hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\mu}}) > 0 \quad (i = 1, \ldots, m).$$

The objective in (2.7) is strictly convex, so $\hat{\rho}$ can be computed by a Newton method (Han, 2014a).

Proposition 1. *The weight $\hat{w}$ in (2.6) satisfies equations (2.4) and (2.5).*

The weight $\hat{w}$ calibrates the fully observed samples so that the weighted average based on the fully observed samples is equal to the unweighted average based on the total samples. In other words, the weighted averages over complete cases match the unweighted averages over the full sample for *all* working propensity and imputation-derived score moments. This multiple balancing is what imparts robustness.

Final MR estimator. We now propose a multiple robust estimator for $\mu(x)$, denoted by $\hat{\mu}_{\mathrm{MR}}$, by re-solving the prediction problem using only complete cases, but weighted by the EL weights:

$$\sum_{i=1}^{m} \hat{w}_i L\left\{Y_i - \mu(X_i)\right\}, \tag{2.8}$$

where the weight $\hat{w}_i$ for every fully observed sample is obtained by (2.6). Intuitively, (2.8) corrects the selection bias in complete cases while pooling information across all working models; if *any one of* π^j or f^k is correctly specified, the calibration in (2.4)–(2.5) aligns the weighted complete-case objective with its full-data counterpart, yielding a consistent estimator.

Remark 1 (Choice of loss)**.** There are various choices of loss function for different tasks. We can choose the least squares loss $L(u) = u^2$ for mean regression of μ_0, or the quantile check loss $L(u) = u\{\tau - \mathbb{1}(u < 0)\}$ for quantile regression of $\mu_\tau(x)$ at the τ-th quantile level.

Remark 2 (Extension to missing covariates)**.** In fact, this framework for missing responses can be easily extended to cases with missing covariates. Without loss of generality, suppose that X_1 contains missing values among $X_1, \cdots, X_p$. In the imputation step (2.2), we draw random samples from the imputation model $f^k(X_1 \mid Y, X_2, \ldots, X_p, b^k)$ and fill the missing values in X_1 with $\hat{X}(\hat{b}^k)$. The objective-function estimates follow steps similar to

those in (2.3) and (2.6); the only difference is that X_i is replaced by the imputed $\hat{X}(\hat{b}^k)$. Finally, we can obtain a multiple robust estimator $\hat{\mu}_{\mathrm{MR}}$ (Han et al., 2019).

2.2 Double Calibration

We now construct a conformal prediction interval framework under missing outcomes by combining split conformal calibration (Vovk et al., 2020) with the multiple-robust (MR) machinery from Section 2.1. The key idea is a *double calibration*: (i) compute conformity scores on a held-out calibration set, and (ii) reweight these scores via empirical-likelihood (EL) balancing so that they represent the full calibration distribution under MAR, even when some working models are misspecified. Without loss of generality, we consider the mean regression estimator for μ_0 throughout this paper. Extensions to quantile estimators and the corresponding theoretical results are shown in S1.5 in the Supplementary Materials.

First, we split the dataset into two subsets, a training set $\mathcal{I}_{tra}$ and a calibration set $\mathcal{I}_{cal}$, with sample sizes n_{tra} and n_{cal}, respectively; the corresponding fully observed sample sizes are denoted by m_{tra} and m_{cal}, respectively. We obtain the multiple robust estimator $\hat{\boldsymbol{\mu}}_{\mathrm{MR}}$ using all training samples $i \in \mathcal{I}_{tra}$ by following the steps in Section 2.1. Let $1-\alpha$ denote the

target coverage rate.

First calibration: conformity scores on complete calibration cases. We conduct the first calibration by calculating the conformity score, defined by the absolute error, using fully observed samples in the calibration set:

$$\hat{\varepsilon}_i = |Y_i - \hat{\mu}_{\mathrm{MR}}(X_i)|\,, \quad i \in \mathcal{I}_{cal} \tag{2.9}$$

and set $m_{\mathrm{cal}} = \sum_{i\in\mathcal{I}_{\mathrm{cal}}} R_i$. The conformity score measures the discrepancy between the prediction and the true response. Let $\mu_0(x) = \mathbb{E}(Y \mid X = x)$ be the unknown population mean regression function. The oracle score is $\varepsilon^* = |Y - \mu_0(X)|$ with $1-\alpha$-quantile $q^*_{1-\alpha}$. Let $\mathbb{P}_{\varepsilon^*}(\cdot)$ denote the distribution function of ε^*; then $\mathbb{P}_{\varepsilon^*}(x < q^*_{1-\alpha}) = 1-\alpha$. A direct approach to approximate $q^*_{1-\alpha}$ is to use empirical quantile estimation of absolute error (2.9). In standard split conformal, one would take the empirical $1-\alpha$-quantile of the fully observed errors $\{\hat{\varepsilon}_i : R_i = 1\}$, denoted by $\hat{q}_{1-\alpha}(\varepsilon_i)$, and form

$$\hat{C}_{\mathrm{CP}}(x) = [\hat{\mu}_{\mathrm{MR}}(X_i) - \hat{q}_{1-\alpha}(\varepsilon_i), \hat{\mu}_{\mathrm{MR}}(X_i) + \hat{q}_{1-\alpha}(\varepsilon_i)], \tag{2.10}$$

where $\hat{q}_{1-\alpha}(\varepsilon_i)$ is the $1-\alpha$-th empirical quantile of the absolute error $\hat{\varepsilon}$ defined in (2.9). However, under MAR the distribution of $\hat{\varepsilon}_i$ among complete cases can differ from that among all calibration points, especially with heteroskedastic or heavy-tailed errors, so $\hat{q}_{1-\alpha}(\hat{\varepsilon})$ may be a biased estimate

of $q^*_{1-\alpha}$. Therefore, we leverage the multiple robust framework to further calibrate the fully observed absolute errors.

Second calibration: MR reweighting of conformity scores. Now we conduct the second calibration of the conformity score using the multiple robust approach. As with outcomes Y_i, conformity scores are missing when $R_i = 0$; we index them by $i = m_{cal} + 1, \ldots, n_{cal}$. We impute their values using the K fitted outcome models from Section 2.1. For each $k = 1, \ldots, K$ and each $i \in \mathcal{I}_{\text{cal}}$ with $R_i = 0$, draw T Monte Carlo imputations

$$Y_i^t(\hat{b}^k) \sim \hat{f}^k(\cdot \mid X_i, \hat{b}^k), \qquad t = 1, \ldots, T,$$

and define the imputed conformity score as $\hat{\varepsilon}_i(\hat{b}^k) = T^{-1} \sum_{i=1}^{T} |Y_i^t(\hat{b}^k) - \hat{\mu}_{\text{MR}(X_i)}|$. Each imputed error $\hat{\varepsilon}_i(\hat{b}^k)$ corresponds to the imputed $Y_i^t(\hat{b}^k)$. Let $\rho_{1-\alpha}(u) = u\{1-\alpha-\mathbb{1}(u < 0)\}$ be the check loss and $\psi_{1-\alpha}(u) = 1-\alpha-\mathbb{1}(u < 0)$ its subgradient at the $1-\alpha$ quantile level. For each model k, obtain a *model-wise* $1-\alpha$-quantile $\hat{q}^{(k)}_{1-\alpha}$ by minimizing the imputed check-loss risk over the entire calibration set,

$$\hat{q}^{(k)}_{1-\alpha} \in \arg\min_{q \in \mathbb{R}} \frac{1}{n_{cal}} \sum_{i \in \mathcal{I}_{\text{cal}}} \Big\{ R_i \, \rho_{1-\alpha}\big(\hat{\varepsilon}_i - q\big) + (1 - R_i) \, \rho_{1-\alpha}\big(\hat{\varepsilon}_i(\hat{b}^k) - q\big) \Big\}, \tag{2.11}$$

Define the corresponding centered ψ-moments by

$$\hat{\xi}^{(k)} \;=\; \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} \Big\{ R_i \, \psi_{1-\alpha}\big(\hat{\varepsilon}_i - \hat{q}^{(k)}_{1-\alpha}\big) + (1-R_i) \, \psi_{1-\alpha}\big(\hat{\varepsilon}_i(\hat{b}^k) - \hat{q}^{(k)}_{1-\alpha}\big) \Big\}. \quad (2.12)$$

In parallel, let $\{\pi^j\}_{j=1}^J$ denote the fitted propensity models defined in Section 2.1 estimated on $\mathcal{I}_{\text{tra}}$, and set

$$\hat{\theta}^{(j)} \;=\; \frac{1}{n_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}}} \pi^j(\hat{a}^j; X_i), \qquad j = 1, \ldots, J. \quad (2.13)$$

We now calibrate the *observed* calibration scores $\{\hat{\varepsilon}_i : R_i = 1\}$ by computing EL weights $\{\hat{d}_i\}_{i \in \mathcal{I}_{\text{cal}},\, R_i = 1}$ that simultaneously balance (i) the propensity moments and (ii) the ψ-moments from (2.12). For each complete calibration case i define the centered moment vector

$$\hat{\boldsymbol{v}}_i(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{q}}) = \big\{ \pi^1\left(\hat{a}^1; X_i\right) - \hat{\theta}^1, \ldots, \pi^J\left(\hat{a}^J; X_i\right) - \hat{\theta}^J, \\ \psi^1\left(\hat{q}^1; X_i\right) - \hat{\xi}^1, \ldots, \psi^K\left(\hat{q}^K; X_i\right) - \hat{\xi}^K \big\}^T,$$

and obtain the EL weight by

$$\hat{d}_i \;=\; \frac{1}{m_{\text{cal}}} \frac{1}{1 + \hat{\boldsymbol{\lambda}}^\top \hat{\boldsymbol{v}}_i}, \qquad \hat{\boldsymbol{\lambda}} = \arg\min_{\lambda} \Big\{ -\frac{1}{m_{\text{cal}}} \sum_{i \in \mathcal{I}_{\text{cal}},\, R_i = 1} \log\left(1 + \hat{\boldsymbol{\lambda}}^\top \hat{\boldsymbol{v}}_i\right) \Big\}, \quad (2.14)$$

with feasibility constraints $1 + \hat{\boldsymbol{\lambda}}^\top \hat{\boldsymbol{v}}_i > 0$ for all $i \in \mathcal{I}_{cal}$ with $R_i = 1$.

Final calibrated quantile and interval. The *double-calibrated* conformity quantile $\hat{q}_{\mathrm{MR}}$ is defined as the root of the weighted score equation

$$\sum_{i \in \mathcal{I}_{\mathrm{cal}},\, R_i=1} \hat{d}_i \, \psi_{1-\alpha}\big(\hat{\varepsilon}_i - \hat{q}_{\mathrm{MR}}\big) \;=\; 0. \tag{2.15}$$

where $\hat{d}_i$ is obtained from (2.14). Therefore, the calibrated quantile estimator for $q^*_{1-\alpha}$ of the conformity scores ε_i is $\hat{q}_{\mathrm{MR}}$.

Finally, we can obtain the conformal prediction interval using multiple robust estimation and double calibration by

$$\hat{C}_{\mathrm{MR}}(x) = \left[\hat{\mu}_{\mathrm{MR}}(x) - \hat{q}_{\mathrm{MR}}, \hat{\mu}_{\mathrm{MR}}(x) + \hat{q}_{\mathrm{MR}}\right]. \tag{2.16}$$

Algorithm 1 in the Appendix summarizes the full **C***onformal Prediction for* **M***issing Data under* **M***ultiple* **R***obust* **L***earning (CM-MRL)* procedure.

Remark 3 (Flexibility in predictors and conformity scores)**.** The double-calibration layer is model-agnostic and offers flexibility in reweighting conformity scores. For instance, the framework allows the point prediction $\hat{\mu}_{\mathrm{MR}}(x)$ in (2.16) to be replaced by predictions from alternative models, including those generated by black-box machine learning techniques. Additionally, the definition of the conformity score, initially presented in (2.9), can be diversified. Alternative formulations of the conformity score, such as those proposed by Romano et al. (2019) and Candès et al. (2023), can

be easily integrated into the proposed framework. We show the details of the extensions in S1.5 in the Supplementary Materials.

3. Theoretical Results

3.1 Marginal and Conditional Coverage

We establish large-sample guarantees for the MR point estimator from Section 2.1 and the double-calibrated quantile from Section 2.2. We first state consistency and asymptotic normality for the proposed estimators. Then, we derive marginal and asymptotic conditional coverage of the conformal interval (2.16). Proof sketches are deferred to S2 in the Supplementary Materials.

(C1). (*i.i.d. sample, MAR, and sample split*) $\{(X_i, R_i, Y_i)\}_{i=1}^{n+1}$ are i.i.d., the split into $\mathcal{I}_{\text{tra}}$ and $\mathcal{I}_{\text{cal}}$ is independent of the data, and $R \perp Y \mid X$ with $0 < \inf_x \pi_0(x) \leq \sup_x \pi_0(x) < 1$.

(C2). (*Oracle error for absolute-residual scores*) Let $\mu_0(x) = \mathbb{E}(Y \mid X = x)$ and $\varepsilon^* = |Y - \mu_0(X)|$. The cdf F_{ε^*} is continuous in a neighborhood of its $1-\alpha$-quantile $q^*_{1-\alpha}$ and has density $f_{\varepsilon^*}(q^*_{1-\alpha}) > 0$.

(C3). (*Modeling robustness*) At least one propensity working model in Π or one outcome model in $\mathcal{F}$ is correctly specified.

(C4). (*Moments complexity*) The fourth moment $\mathbb{E}\|X\|^4 < \infty$, and the classes $\{\pi^j(\cdot; a)\}$, $\{f^k(\cdot \mid \cdot; b)\}$ have manageable complexity so that the empirical processes used in calibration are $o_p(1)$, i.e., Donsker classes.

(C5). (*Loss regularity*) The loss $L(\cdot)$ is convex and locally Lipschitz; for mean regression $L(u) = u^2$, for τ-quantiles $L(u) = u\{\tau - \mathbb{1}(u < 0)\}$ with subgradient $\psi_\tau(u) = \tau - \mathbb{1}(u < 0)$.

Condition (C1) gives the basic independence assumptions required for conformal prediction. Condition (C2) allows heteroskedasticity in $Y \mid X$; it only requires smoothness of the *marginal* distribution of ε^*. If one prefers a conditional formulation, it suffices to assume that the conditional cdf of $|Y - \mu_0(X)|$ given X has a density that is bounded and bounded away from zero at the τ-quantile uniformly in x. Condition (C3) is a regularity assumption for quantile estimation. Condition (C4) requires $E\|X\|^4$ to be bounded. It controls the tail behavior of the distribution of X and is used to establish a Donsker class in empirical-process arguments.

Remark 4. Assumption (C4) holds in various standard settings. For instance, if X has bounded support, X is sub-Gaussian, or X is categorical, then $\mathbb{E}\|X\|^4 < \infty$. Moreover, when the propensity score model $\pi(\cdot)$ and the

working outcome models $f(\cdot)$ are finite smooth functions such as logistic π and Gaussian linear models for $Y \mid X$ with parameters in compact sets, the corresponding function classes satisfy (C4).

Remark 5. Assumption (C4) is a high-level empirical-process condition ensuring asymptotic normality of the plug-in estimating equations. Following Chernozhukov et al. (2018), one can replace the Donsker and moment conditions of (C4) by employing cross-fitting and Neyman-orthogonal scores. An extension of this argument to the full CM–MRL calibration moments is shown in the Supplementary Materials S1.6.

Theorem 1 (Consistency and asymptotic normality of MR estimator)**.** *Under (C1)–(C5), if either a propensity model in* Π *or an outcome model in* $\mathcal{F}$ *is correctly specified, then* $\hat{\mu}_{\mathrm{MR}} \xrightarrow{p} \mu_0$. *If, in addition,* $f_{\varepsilon|X}(0 \mid x)$ *exists and is bounded away from zero uniformly in* x*, then* $\sqrt{n}\,(\hat{\mu}_{\mathrm{MR}} - \mu_0) \to \mathcal{N}\big(0,\, V_{\mathrm{MR}}\big)$*, where* V_{MR} *is the semiparametric variance bound when a propensity model and an outcome model are both correctly specified.*

The variance V_{MR} and the proof are given in Han, 2014b; Han et al., 2019. This result demonstrates the consistency of the mean regression estimator $\hat{\mu}_{\mathrm{MR}}$. Additionally, the asymptotic normality of the estimator holds under further assumptions. We now show the consistency of the double-

calibrated estimator $\hat{q}_{MR}$ for the reweighted conformity score in Section 2.2.

Theorem 2 (Consistency and limit distribution of $\hat{q}_{\mathrm{MR}}$)**.** *Under (C1)–(C5),* $\hat{q}_{\mathrm{MR}} \xrightarrow{p} q^*_{1-\alpha}$. *Moreover, if* f_{ε^*} *is continuously differentiable at* $q^*_{1-\alpha}$*, then, with* $m_{\mathrm{cal}} = \sum_{i \in \mathcal{I}_{\mathrm{cal}}} R_i$,

$$\sqrt{m_{\mathrm{cal}}}\left(\hat{q}_{\mathrm{MR}} - q^*_{1-\alpha}\right) \to \mathcal{N}\Big(0,\ \frac{\alpha(1-\alpha)}{\{f_{\varepsilon^*}(q^*_{1-\alpha})\}^2} \cdot \kappa_{\mathrm{MR}}\Big),$$

where $\kappa_{\mathrm{MR}} \leq 1$ *is an efficiency factor determined by the EL calibration;* $\kappa_{\mathrm{MR}} = 1$ *when weights are uniform in complete-case split conformal and typically* $\kappa_{\mathrm{MR}} < 1$ *when the balancing moments are correctly specified.*

Even when the point estimator is misspecified, the EL calibration aligns the complete-case score distribution with the full calibration distribution by constraints (2.4) and (2.5). As a result, $\hat{q}_{\mathrm{MR}}$ converges to $q^*_{1-\alpha}$ and has smaller asymptotic variance than the unweighted empirical quantile when the balancing moments are correct. This typically yields *shorter* intervals than complete-case split conformal while preserving coverage.

Remark 6. Although Theorems 1 and 2 and the related conditions are stated for mean regression with squared loss, CM–MRL can be extended directly to conditional quantile estimation $\mu_\tau(x) = Q_\tau(Y \mid X = x)$, yielding MR estimators of $\hat{f}_{\alpha/2}$ and $\hat{f}_{1-\alpha/2}$ as shown in Han et al. (2019). In addition,

we can extend the method to the CQR conformity score (Romano et al., 2019) by applying the same EL reweighting to complete calibration scores and targeting the $(1-\alpha)$-quantile of the corresponding oracle score $S^* = \max\{Y - f_{0,1-\alpha/2}(X),\ f_{0,\alpha/2}(X) - Y\}$. The extended asymptotic results and proof details are shown in S1.5 in the Supplementary Materials.

The conformal interval constructed in (2.16) has strong theoretical support, providing reliable coverage both marginally and conditionally.

Theorem 3 (Asymptotic marginal coverage)**.** *Under (C1)–(C5),*

$$\liminf_{n\to\infty}\ \mathbb{P}\big\{\, Y_{n+1} \in \hat{C}_{\mathrm{MR}}(X_{n+1}) \,\big\}\ \geq\ 1-\alpha.$$

Additionally, if the calibration weights are functions of X only and the split is independent, then the coverage error is $o(1)$.

The result states that double calibration recovers the oracle conformity quantile and thus achieves the nominal marginal coverage asymptotically.

To quantify the accuracy of the proposed interval, we show the conditional coverage. There are various metrics for measuring conditional coverage (Lei et al., 2018; Feldman et al., 2021; Foygel Barber et al., 2021). We adopt the asymptotic conditional coverage definition in Sesia and Romano (2021), where conditional coverage holds when intervals are asymptotically

close to the oracle intervals under regular conditions. We define the oracle interval by the true conditional mean $\mu_0(x)$ and quantile $q^*_{1-\alpha}$.

Definition 1 (Oracle prediction interval)**.** Let $C^*(x) = \big[\mu_0(x) - q^*_{1-\alpha},\ \mu_0(x) + q^*_{1-\alpha}\big]$ denote the oracle interval.

Theorem 4 (Asymptotic conditional coverage)**.** *Under (C1)–(C5), there exist sequences $\xi_n \to 0$ and $b_n \to 0$ such that*

$$\mathbb{P}\Big\{ \mathbb{P}\big[Y_{n+1} \in \hat{C}_{\mathrm{MR}}(X_{n+1}) \;\big|\; X_{n+1} \big] \;\geq\; 1 - \alpha - \xi_n \Big\} \;\longrightarrow\; 1 - b_n.$$

Equivalently, $\hat{C}_{\mathrm{MR}}(\cdot)$ is asymptotically equivalent to $C^*(\cdot)$.

4. Numerical Results

4.1 Simulation Study

In this section, we evaluate the performance of the proposed CM–MRL via simulations and compare it with various existing methods under different scenarios and settings. We provide two experiments: first, we consider a setting adapted from Han (2014b) to assess effectiveness and robustness, and then we compare against related conformal methods.

4.1.1 Numerical experiment I

We generate i.i.d. covariates by

$$X_1 \sim \mathcal{N}(5,1), \quad X_2 \sim \text{Bernoulli}(0.5), \quad X_3 \sim \mathcal{N}(0,1), \quad X_4 \sim \mathcal{N}(0,1).$$

The true regression is linear with $\beta_0 = (3.5, 0.5, 2.0, 1.0, 1.0)$. The outcome follows a linear signal-plus-noise model

$$Y \;=\; 3.5 + 0.5X_1 + 2X_2 + X_3 + X_4 + \sigma\,\varepsilon_Y,$$

so that the conditional mean is $\mu_0(x) = \beta_0^\top x$ with $\beta_0 = (3.5,\ 0.5,\ 2,\ 1,\ 1)$. We generate MAR missingness via $\text{logit}\,\pi_0(X) = 3.5 - 5.0X_2$ with about 58% observed and consider three error scenarios, including Gaussian, heavy-tailed, and heteroskedastic errors. We evaluate robustness under misspecification using two candidate propensity models and two candidate outcome models across the six settings shown in Table 2; at least one candidate model is correct in each setting. Importantly, both candidate propensity score models are correctly specified in Experiment I: the first is the true logistic model, and the second is a larger correctly specified logistic model that contains the true propensity model as a special case. We run 100 Monte Carlo replications with $n = 3000$, split the sample into training/calibration/testing, and compare CM–MRL with Impute–SC, complete-case split conformal (SC), and WCCQR/WCCQR–CV at target coverage

$1-\alpha = 0.9$; full details are provided in S1.2 in the Supplementary Materials.

Performance is summarized by empirical coverage, average interval length, and the corresponding standard deviation (sd) in Tables 3 and 4. In addition, Figures 1 and 2 report boxplots of the length across all methods under different settings and scenarios.

Across all settings, complete-case split conformal (SC) systematically overcovers the nominal level $1 - \alpha = 0.9$ and produces the longest intervals. In contrast, the three missing-data methods, **CM–MRL** (ours), Impute–SC, and WCCQR/WCCQR–CV, cluster near nominal coverage with markedly shorter lengths. In settings S1 and S2 with correct $a^{(1)}$ available, CM–MRL, Impute–SC, and WCCQR/WCCQR–CV achieve comparable coverage around 0.89–0.90 with similar lengths, while SC is longer and overconservative. For example, in S1A, CM–MRL length is 4.621 versus 5.133 for SC; in S2C, CM–MRL is 4.344 versus 4.878 for SC. Standard deviations are small and similar across the missing-data methods. In Setting S3 with no correct outcome model, CM–MRL delivers the shortest intervals among all methods, reflecting its calibration efficiency. SC again overcovers and is longest. This highlights a trade-off under misspecification: CM–MRL prioritizes shorter length while maintaining stable performance, whereas imputation/weighting baselines tilt toward slightly higher coverage

with wider intervals.

Therefore, CM–MRL consistently produces intervals that are *(i)* much shorter than complete-case SC across scenarios and *(ii)* competitive with, or shorter than, Impute–SC/WCCQR when a correct working model is present (S1/S2). Under outcome-model misspecification (S3/S4), and when only the larger propensity model is used (S5/S6), CM–MRL preserves its short-length advantage but exhibits mild undercoverage, illustrating the expected robustness and efficiency trade-off. In low-noise regimes, such as scenario (A), it offers the best length-coverage balance among all missing-data methods.

4.1.2 Numerical experiment II

In the second experiment, we consider a nonlinear regression model with $X \in \mathbb{R}^6$ and heteroskedastic noise. We generate $n = 2{,}000$ i.i.d. observations and repeat the experiment 100 times. We conduct a nonlinear, heteroskedastic regression simulation with MAR missing outcomes under three propensity scenarios (A–C), using a train/calibration/test split (50%/30%/20%), and compare CM–MRL with Han-Bootstrap and weighted conformal baselines (CP-Logit/CP-NN/CP-RF). Full data-generating details and model specifications are in S1.2 in the Supplementary Materials.

The performance of each method is shown in Figure 3, which displays coverage in blue bars associated with the left axis and mean interval length in orange bars associated with the right axis, with the dashed horizontal line indicating the nominal $1 - \alpha = 0.90$ coverage. The plots show a clear coverage-length balance across intervals and scenarios. Across the three MAR scenarios, CM–MRL consistently delivers among the shortest mean interval lengths while keeping coverage close to the 0.90 target, especially in Scenarios A and B. In contrast, Han-Bootstrap produces substantially wider intervals. The weighted conformal baselines CP-Logit and CP-RF often achieve coverage above the target, but their intervals inflate noticeably as the missingness model becomes more nonlinear and overlap weakens in Scenario C, which is consistent with instability from IPW weights under harder MAR with larger weights. Finally, CP-NN produces very short intervals, but its coverage drops below 0.90 in these plots, making it anti-conservative. This is a common failure mode when flexible propensity models are trained on finite samples, leading to underestimated calibration quantiles and undercoverage. Therefore, CM–MRL is more efficient and more robust across MAR complexities.

Table 1: Choices of the error term ε_Y and the scale σ.

No.	σ	ε_Y
A	1	$\varepsilon_Y \sim \mathcal{N}(0,1)$
B	1	$\varepsilon_Y \sim t_3$
C	0.6	$\varepsilon_Y \sim \mathcal{N}\big(0,\,(0.6+0.2\lvert X_1\rvert)^2\big)$

Note. We use all three choices in every setting. t_ν denotes a Student-t distribution with ν degrees of freedom. The probability of outlier occurrence is 0.02 in choice C.

Table 2: Candidate propensity and imputation model combinations.

Setting	$\pi^1(\boldsymbol{\alpha}^1)$	$\pi^2(\boldsymbol{\alpha}^2)$	$a^1(\boldsymbol{\gamma}^1)$	$a^2(\boldsymbol{\gamma}^2)$
S1	✓	✓	✓	✓
S2	✓	✗	✓	✓
S3	✓	✓	✗	✓
S4	✓	✓	✓	✗
S5	✗	✓	✓	✓
S6	✗	✓	✓	✗

Note. ✓ indicates the model is adopted, while ✗ indicates not adopted.

4.2 Real Data

As an application of the proposed method, we consider data from 2,139 HIV-infected subjects enrolled in the AIDS Clinical Trials Group Protocol 175 (ACTG 175) (Hammer et al., 1996). Subjects were randomly

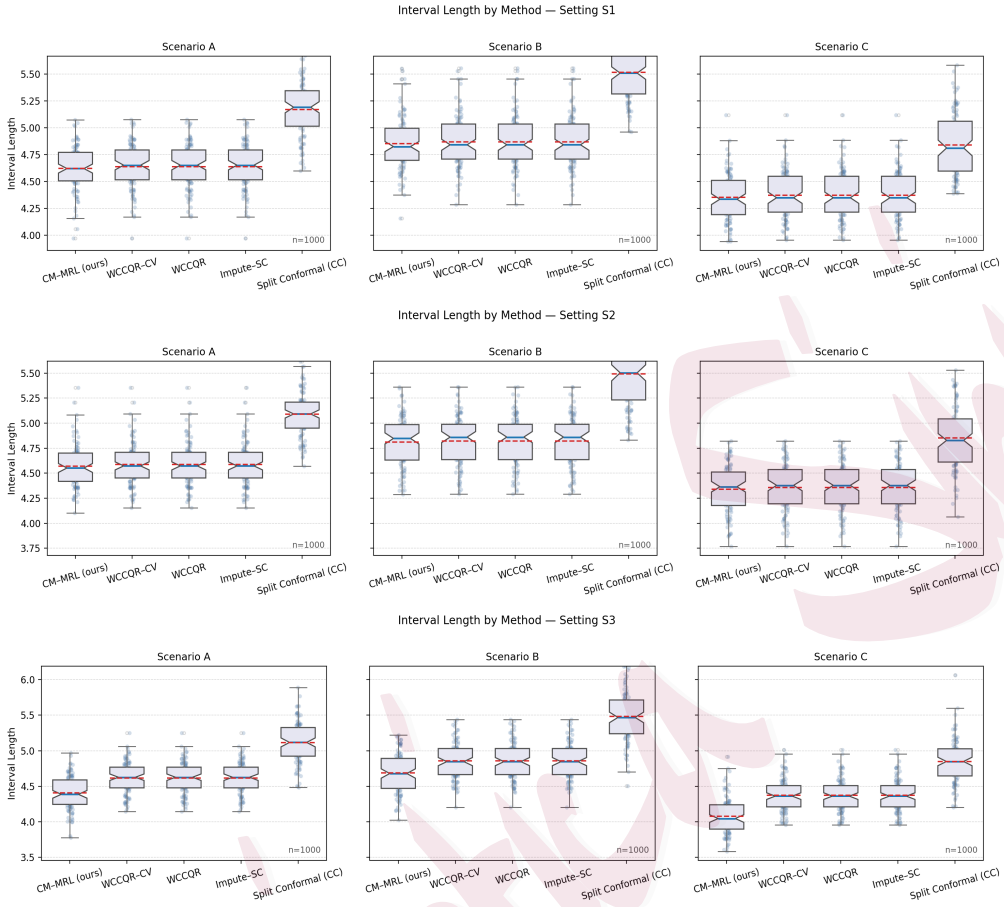

Figure 1: Interval length of all methods in Settings S1-S3 with candidate model selections under Scenarios A-C with 100 repetitions.

assigned to one of four antiretroviral regimens: zidovudine (ZDV) alone, ZDV+didanosine (ddI), ZDV+zalcitabine (ddC), or ddI alone.

Following the analyses by Han (2014b), we compare two treatment arms: the standard ZDV monotherapy arm and the combined arm of the three newer treatments. These two arms include 532 and 1607 subjects,

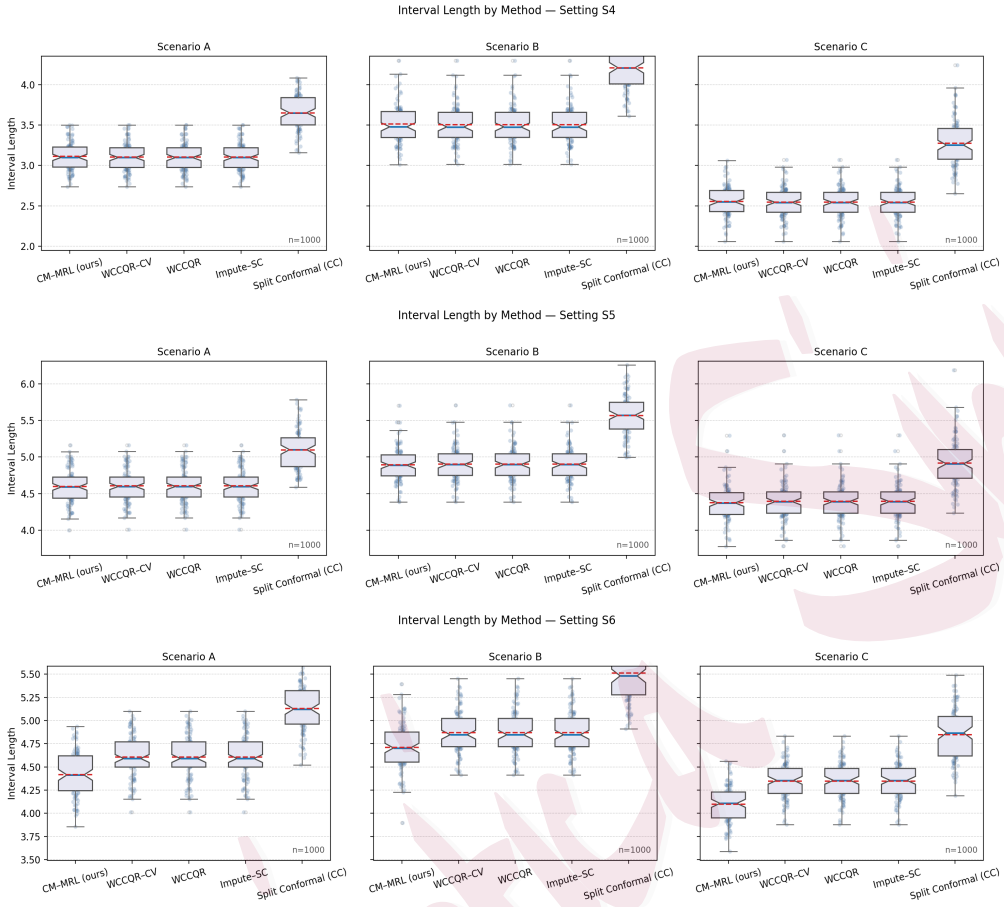

Figure 2: Interval length of all methods in Settings S4-S6 with candidate model selections under Scenarios A-C with 100 repetitions.

respectively. Our primary focus is the effect of the treatment arm on CD4 counts measured at 96 ± 5 weeks post-baseline ($CD4_{96}$), adjusting for baseline CD4 counts ($CD4_0$) and other baseline characteristics. These character-

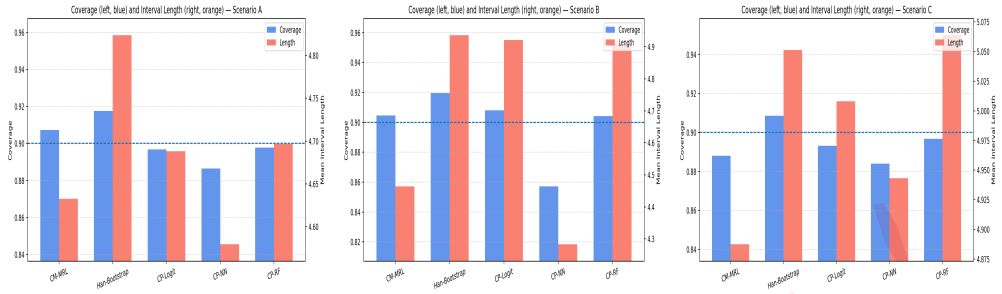

Figure 3: Average interval length (orange bar) and coverage (blue bar) in numerical experiment II under Scenarios A-C over 100 repetitions.

istics include continuous covariates (age in years and weight in kilograms) and binary covariates (treatment, where 0 = ZDV; race, where 0 = white; gender, where 0 = female; antiretroviral history, where 0 = naive and 1 = experienced; and whether the subject was off-treatment prior to 96 weeks, where 0 = no). Our aim is to fit the following linear regression model:

$$\begin{aligned} \text{CD}_{96} =& \beta_1 + \beta_2 \text{ trt } + \beta_3 \text{CD}_0 + \beta_4 \text{ age } + \beta_5 \text{ weight } + \beta_6 \text{ race} \\ &+ \beta_7 \text{ gender } + \beta_8 \text{ history } + \beta_9 \text{ offtrt } + \epsilon, \end{aligned}$$

where ϵ has a mean of zero conditional on all covariates. The data can be accessed through the R package *speff2trial*. We provide summary statistics for these data in Table 1.

Accurately specifying a model for $\pi(\boldsymbol{X})$ is challenging with an eight-

http://cran.r-project.org/web/packages/speff2trial/speff2trial.pdf

dimensional $\boldsymbol{X}$, even with model selection and diagnostic techniques. The same challenge applies to modeling $\mathbb{E}(Y \mid \boldsymbol{X})$. Due to potential model misspecifications, the reliability of estimation and inference based on doubly robust methods can be questionable. Therefore, we apply the proposed method, which, although not definitive, may provide more reliable conclusions in the presence of model misspecifications, as demonstrated by the simulation studies in Section 4.1. Following Han (2014b), we use a logistic regression model for the propensity score function $\pi(\boldsymbol{X})$ and a linear regression model for $\mathbb{E}(Y \mid \boldsymbol{X})$. To ensure thoroughness, both models include all main effects of $\boldsymbol{X}$. We employ the estimating function $\boldsymbol{U}(Y, \boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{X}\left(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}\right)$.

We further compare CM–MRL with split conformal and weighted conformal baselines whose propensity scores are estimated using machine-learning methods, such as neural networks (CP-NN) and random forests (CP-RF), implemented via the `SuperLearner` package in R; see S1.3 in Supplementary Materials for full implementation details and settings.

Table 5 reports prediction-interval performance on the ACTG175 test set. Overall, the compared methods yield coverages ranging from 0.871 to 0.899 at the nominal $1 - \alpha = 0.90$ level. Split conformal attains the

https://cran.r-project.org/web/packages/SuperLearner/index.html

coverage closest to nominal (0.899) with mean length 170.35. CM–MRL achieves comparable coverage (0.889) while producing a shorter interval (166.71), improving efficiency relative to split conformal with only a modest reduction in coverage. Among the weighted conformal baselines, CP–RF has a similar mean length to CM–MRL (166.72) but slightly lower coverage (0.887), whereas CP–NN yields the shortest intervals (162.79) but exhibits the largest undercoverage (0.871). Taken together, these results suggest that CM–MRL provides a good validity–efficiency trade-off in this application. It maintains near-nominal coverage while delivering comparatively parsimonious intervals, and it is less sensitive to the choice of a single propensity model by leveraging multiple-robust calibration.

Supplementary Materials

In the Supplementary Materials, we present extensions of the proposed method, including the algorithm pseudocode (S1.1), detailed experiment settings (S1.2, S1.3), extensions under model misspecification (S1.4), the quantile conformity score version (S1.5), and the double machine learning extension (S1.6). We also include technical proofs of the main results (S2).

References

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.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Feldman, S., S. Bates, and Y. Romano (2021). Improving conditional coverage via orthogonal quantile regression. *Advances in neural information processing systems 34*, 2060–2071.

Foygel Barber, R., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA 10*(2), 455–482.

Gibbs, I. and E. Candes (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems 34*, 1660–1672.

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.

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference 148*, 101–110.

Han, P. (2014b). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association 109*(507), 1159–1173.

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika 103*(3), 683–700.

Han, P., L. Kong, J. Zhao, and X. Zhou (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 81*(2), 305–333.

Han, P. and L. Wang (2013). Estimation with missing data: beyond double robustness. *Biometrika 100*(2), 417–430.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.

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., J. Robins, and L. Wasserman (2013). Distribution-free prediction sets. *Journal of the*

American Statistical Association 108(501), 278–287.

Lei, J. and L. Wasserman (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 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, W., Y. Gu, and L. Liu (2020). Demystifying a class of multiply robust estimators. *Biometrika 107*(4), 919–933.

Little, R. J. and D. B. Rubin (2002). Maximum likelihood for general patterns of missing data: Introduction and theory with ignorable nonresponse. *Statistical analysis with missing data*, 164–189.

Little, R. J. and D. B. Rubin (2019). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Romano, Y., E. Patterson, and E. Candes (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems 32*, 3543–3553.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observa-

tional studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association 91*(434), 473–489.

Rubin, D. B. (2018). Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pp. 29–62. Chapman and Hall/CRC.

Sesia, M. and Y. Romano (2021). Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems 34*, 6304–6315.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika 97*(3), 661–682.

Tibshirani, R. J., R. Foygel Barber, E. Candes, and A. Ramdas (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems 32*.

Tu, R., C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang (2019). Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1762–1770. PMLR.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Vovk, V., I. Petej, P. Toccaceli, A. Gammerman, E. Ahlberg, and L. Carlsson (2020). Conformal

calibrators. In *conformal and probabilistic prediction and applications*, pp. 84–99. PMLR.

Zaffran, M., A. Dieuleveut, J. Josse, and Y. Romano (2023). Conformal prediction with missing values. *arXiv preprint arXiv:2306.02732*.

Zaffran, M., O. Féron, Y. Goude, J. Josse, and A. Dieuleveut (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR.

Wenlu Tang: University of Alberta

wenlu4@ualberta.ca https://orcid.org/0000-0001-9448-578X

Hongni Wang: Shandong University of Finance and Economics

wanghongnisd@126.com

Xingcai Zhou: Nanjing Audit University

xczhoustat@126.com,

Bei Jiang: University of Alberta

bei1@ualberta.ca https://orcid.org/0000-0002-0033-839X

Linglong Kong: University of Alberta

lkong@ualberta.ca https://orcid.org/0000-0003-3011-9216

Table 3: Coverage and interval length (mean (sd)) across methods in Settings S1-S4 under 100 repetitions.

Method	A		B		C	
	Cov.	**Len.**	**Cov.**	**Len.**	**Cov.**	**Len.**
Setting S1						
CM–MRL	0.892(0.019)	4.620(0.213)	0.885(0.021)	4.849(0.257)	0.898(0.019)	4.362(0.248)
Impute–SC	0.893(0.018)	4.638(0.212)	0.887(0.019)	4.868(0.246)	0.899(0.018)	4.386(0.246)
Split Conformal (CC)	0.927(0.016)	5.169(0.268)	0.923(0.017)	5.547(0.300)	0.932(0.018)	4.936(0.319)
WCCQR	0.893(0.018)	4.638(0.212)	0.887(0.019)	4.868(0.246)	0.899(0.018)	4.386(0.246)
WCCQR–CV	0.893(0.018)	4.638(0.212)	0.887(0.019)	4.868(0.246)	0.899(0.018)	4.386(0.246)
Setting S2						
CM–MRL	0.887(0.020)	4.558(0.243)	0.886(0.019)	4.819(0.248)	0.896(0.018)	4.341(0.222)
Impute–SC	0.888(0.018)	4.571(0.235)	0.888(0.018)	4.833(0.234)	0.897(0.018)	4.356(0.221)
Split Conformal (CC)	0.924(0.015)	5.072(0.253)	0.925(0.017)	5.498(0.282)	0.930(0.017)	4.915(0.277)
WCCQR	0.888(0.018)	4.571(0.235)	0.888(0.018)	4.833(0.234)	0.897(0.018)	4.356(0.221)
WCCQR–CV	0.888(0.018)	4.571(0.235)	0.888(0.018)	4.833(0.234)	0.897(0.018)	4.356(0.221)
Setting S3						
CM–MRL	0.871(0.025)	4.395(0.284)	0.872(0.022)	4.620(0.285)	0.875(0.023)	4.077(0.261)
Impute–SC	0.892(0.020)	4.639(0.233)	0.885(0.021)	4.816(0.255)	0.898(0.020)	4.396(0.229)
Split Conformal (CC)	0.925(0.017)	5.176(0.290)	0.922(0.017)	5.432(0.316)	0.932(0.017)	4.943(0.313)
WCCQR	0.892(0.020)	4.639(0.233)	0.885(0.021)	4.816(0.255)	0.898(0.020)	4.396(0.229)
WCCQR–CV	0.892(0.020)	4.639(0.233)	0.885(0.021)	4.816(0.255)	0.898(0.020)	4.396(0.229)
Setting S4						
CM–MRL	0.879(0.021)	3.134(0.160)	0.872(0.023)	3.487(0.201)	0.877(0.021)	2.504(0.175)
Impute–SC	0.879(0.021)	3.133(0.160)	0.872(0.023)	3.487(0.201)	0.875(0.019)	2.495(0.171)
Split Conformal (CC)	0.930(0.015)	3.710(0.207)	0.924(0.017)	4.224(0.266)	0.936(0.015)	3.344(0.262)
WCCQR	0.879(0.021)	3.133(0.160)	0.872(0.023)	3.487(0.201)	0.875(0.019)	2.495(0.171)
WCCQR–CV	0.879(0.021)	3.133(0.160)	0.872(0.023)	3.487(0.201)	0.875(0.019)	2.495(0.171)

Table 4: Coverage and interval length (mean (sd)) across methods in Settings S5-S6 under 100 repetitions.

Method	A		B		C	
	Cov.	**Len.**	**Cov.**	**Len.**	**Cov.**	**Len.**
Setting S5						
CM–MRL	0.880(0.023)	4.464(0.251)	0.877(0.023)	4.707(0.261)	0.885(0.021)	4.169(0.238)
Impute–SC	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)
Split Conformal (CC)	0.925(0.017)	5.175(0.291)	0.922(0.017)	5.433(0.316)	0.932(0.017)	4.944(0.313)
WCCQR	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)
WCCQR–CV	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)
Setting S6						
CM–MRL	0.865(0.026)	4.299(0.282)	0.867(0.024)	4.538(0.295)	0.871(0.025)	3.984(0.273)
Impute–SC	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)
Split Conformal (CC)	0.925(0.017)	5.175(0.291)	0.922(0.017)	5.433(0.316)	0.932(0.017)	4.944(0.313)
WCCQR	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)
WCCQR–CV	0.892(0.020)	4.640(0.233)	0.884(0.021)	4.816(0.256)	0.898(0.019)	4.398(0.230)

Table 5: Prediction-interval performance on the ACTG175 test set.

Method	**Mean length**	**Coverage**	$n_{\textbf{obs test}}$
CM–MRL	166.71	0.889	280
Split conformal	170.35	0.899	280
CP–RF	166.72	0.887	280
CP–NN	162.79	0.871	280