Debiased Machine Learning for Conformal Prediction of Counterfactual Outcomes Under Runtime Confounding

Keith Barnatchez KEITHBARNATCHEZ@G.HARVARD.EDU
Department of Biostatistics, Harvard T.H. Chan School of Public Health

Kevin P. Josey KEVIN.JOSEY@CUANSCHUTZ.EDU
Department of Biostatistics and Informatics, Colorado School of Public Health

Rachel C. Nethery RNETHERY@HSPH.HARVARD.EDU
Department of Biostatistics, Harvard T.H. Chan School of Public Health

Giovanni Parmigiani GP@JIMMY.HARVARD.EDU
Department of Data Science, Dana Farber Cancer Institute

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Data-driven decision making frequently relies on predicting counterfactual outcomes. In practice, researchers commonly train counterfactual prediction models on a source dataset to inform decisions on a possibly separate target population. Conformal prediction has arisen as a popular method for producing assumption-lean prediction intervals for counterfactual outcomes that would arise under different treatment decisions in the target population of interest. However, existing methods require that every confounding factor of the treatment-outcome relationship used for training on the source data is additionally measured in the target population, risking miscoverage if important confounders are unmeasured in the target population. In this paper, we introduce a computationally efficient debiased machine learning framework that allows for valid prediction intervals when only a subset of confounders is measured in the target population, a common challenge referred to as runtime confounding. Grounded in semiparametric efficiency theory, we show the resulting prediction intervals achieve desired coverage rates with faster convergence compared to standard methods. Through numerous synthetic and semi-synthetic experiments, we demonstrate the utility of our proposed method.

Keywords: Conformal prediction, Counterfactual prediction, Debiased machine learning, Runtime confounding, Influence curve

1. Introduction

Data-driven decision-support tools (DSTs) are experiencing rapid growth across diverse domains, including personalized medicine, marketing, and social services (Musen et al., 2021; Chouldechova et al., 2018; Fischer-Abaigar et al., 2024). Of particular value are DSTs which predict individual-level *counterfactual outcomes* arising from different possible actions, or treatments, performed by the decision-maker. Recognizing that decision-makers often require *uncertainty quantification* around individual-level counterfactual predictions, an emerging interdisciplinary literature has developed at the intersection of causal inference, statistics and machine learning focused on constructing robust prediction intervals for counterfactual outcomes predicted from flexible, nonparametric models.

Beginning with the seminal work of Lei and Candès (2021), there has been a growing effort to extend tools from conformal prediction to enable the formation of prediction intervals of counterfactual outcomes and individual treatment effects (Yang et al., 2024; Liu et al., 2024; Alaa et al., 2023; Gao et al., 2025; Schröder et al., 2025). Despite the large interest in methods for constructing robust prediction intervals in settings ranging from covariate shift to surrogate outcomes, numerous practical challenges remain unaddressed. One particular challenge is *runtime confounding*, where only a subset of the covariates needed to adjust for confounding of the treatment-outcome relationship are measured on units in the target population where counterfactual predictions are desired (Coston et al., 2020).

Runtime confounding arises frequently in practice, and is typically induced when one is able to collect extensive covariate information for training in a source population, but collecting information on this full set of covariates in the target population of interest is cost prohibitive or infeasible. For instance, in personalized medicine, electronic health records may contain detailed patient histories during model training, but point-of-care decisions often rely on limited covariate measurements (Deschepper et al., 2025; Collins and Dhiman, 2023). Similarly, in marketing applications, customer profiles built from historical data may be unavailable due to privacy regulations when making real-time recommendations (Bleier et al., 2020). Naively restricting models to train on only the set of confounders available in both the source and target populations risks yielding inaccurate prediction intervals when discarded variables serve as confounders of the treatment-outcome relationship. Despite this, existing methods typically assume full access to confounders in the source and target populations, leaving researchers with little recourse to address this challenge.

Contributions. In this work, we propose methods for performing conformal prediction of counterfactual outcomes which address the critical challenge of runtime confounding. Our proposed approach leverages tools from semiparametric theory and debiased machine learning (DML) (Park and Cho, 2025), ensuring constructed intervals attain valid coverage under a modest set of conditions relative to competing methods. We provide a computationally efficient implementation of our approach, which avoids challenges commonly faced by DML methods. Through numerous numerical experiments, we compare our proposed method to alternative approaches based on existing popular frameworks, demonstrating conditions under which our method tends to outperform the latter methods. We additionally derive valid loss functions for counterfactual quantile regression under runtime confounding settings, where the resulting predictions can be used to construct quantile conformity scores (Romano et al., 2019) within our proposed framework. Although not our primary contribution, we additionally provide a weighted conformal prediction method capable of addressing runtime confounding.

Related work. Our work is situated at the intersection of causal inference, conformal prediction and transfer learning. Leveraging results from Tibshirani et al. (2019) and Romano et al. (2019), Lei and Candès (2021) introduced weighted quantile conformal prediction to construct intervals for counterfactuals, addressing covariate shift across treatment levels. Subsequent work has extended these ideas, implementing doubly-robust methods addressing covariate shift and multi-study settings (Yang et al., 2024; Liu et al., 2024), accounting for surrogate outcomes (Gao et al., 2025), and basing scores on meta-learners of counterfactual outcomes and treatment (Alaa et al., 2023). Our work explicitly addresses covariate shift across treatment levels within the source population, as well as between target and source populations, while allowing for incomplete confounder information in the target population.

The causal transfer learning literature aims to address distribution shifts between source and target population data in order to estimate marginal and conditional causal effects (Shyr et al., 2025; Bica and van der Schaar, 2022; Colnet et al., 2024; Rojas-Carulla et al., 2018; Voter et al., 2025). Recent work has developed doubly-robust methods for unknown shifts based in semiparametric theory (Graham et al., 2024; Zeng et al., 2025), with numerous extensions accommodating multi-source data and privacy constraints (Han et al., 2025, 2023) and the incorporation of surrogate outcomes (Kallus and Mao, 2025). We contribute to this literature by deriving valid loss functions for causal quantile regression that enable learning target population counterfactual quantile functions, accounting for covariate shift across treatment levels and populations.

The problem of runtime confounding for point prediction of counterfactual outcomes was formalized by Coston et al. (2020). Our work is the first to extend this problem to the construction of *prediction intervals* through semiparametric efficient conformal prediction. By allowing for covariate shift across the target and source populations we extend the work of Coston et al. (2020), who only considered covariate shift across treatment levels within the source population. More broadly, our setting is connected to a wider literature on counterfactual prediction in which the deployed prediction rule may condition on only a subset of confounders, including work motivated by transportability and time-varying covariate information (Boyer et al., 2025; Keogh and Van Geloven, 2024).

2. Problem Setting and Background

We consider a setting where researchers are interested in the relationship between a categorical treatment variable A taking on values in a set $\mathcal{A}$ and an outcome of interest Y. Complete information on Y and A is provided for all units in a source population dataset, while both Y and A are unavailable in a target population dataset, consistent with a setting where target population members have not yet received treatment. Source population membership is denoted by the indicator variable S. It is assumed there is a set of baseline covariates $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{U})$ that are fully measured in the source population dataset, whereas only a subset of these covariates, $\boldsymbol{V}$, are measured in the target population. The induced data structure can be characterized by the observational unit,

$$\boldsymbol{O}_i = (S_i Y_i,\ S_i A_i,\ S_i \boldsymbol{U}_i,\ \boldsymbol{V}_i,\ S_i) \sim \mathbb{P}, \quad i = 1, \ldots, n,$$

where we adopt the convention that for any random variable Z, its observed value is $SZ + (1-S)\texttt{NA}$ to make explicit that Y, A and $\boldsymbol{U}$ are only observed when the source data indicator $S = 1$. Following Coston et al. (2020), we refer to this setting as **runtime confounding**, since $\boldsymbol{U}$ may contain potential confounding factors of the relationship between A and Y.

Table 1 further summarizes this data structure. We note that such a data structure could arise in both single- or multi-source settings. For instance, the above data structure could arise from a setting where an initial batch of data on Y, A and $\boldsymbol{X}$ is collected from a single site, and additional observations on only $\boldsymbol{V}$ are collected by the same site, possibly according to a different sampling strategy that induces covariate shift. Similarly, the above structure could arise if the target population observations $\boldsymbol{V}$ are collected at an external site which lacks the capacity to measure the additional confounding variables $\boldsymbol{U}$, but wishes to use models trained in the source population to construct prediction intervals for its units.

Objective: We fix interest on the construction of prediction intervals for counterfactual outcomes of subjects in the target population. Following the Rubin potential outcomes framework (Little

Y	A	$\boldsymbol{V}$	$\boldsymbol{U}$	S
Y_1	A_1	$\boldsymbol{V}_1$	$\boldsymbol{U}_1$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{n_1}	A_{n_1}	$\boldsymbol{V}_{n_1}$	$\boldsymbol{U}_{n_1}$	1
NA	NA	$\boldsymbol{V}_{n_1+1}$	NA	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
NA	NA	$\boldsymbol{V}_{n_1+n_0}$	NA	0

Table 1: Observed data structure.

and Rubin, 2000), we let $Y_i(a)$ denote the counterfactual outcome that subject i would experience under treatment level $A_i = a$. Our primary goal is to form prediction intervals $C_a(\boldsymbol{V})$ that cover counterfactual outcomes in the target population with a desired coverage probability:

$$\mathbb{P}(Y(a) \in C_a(\boldsymbol{V})|S = 0) \geq 1 - \alpha, \;\; a \in \mathcal{A},$$

where $C_a(\boldsymbol{V})$ is a function of $\boldsymbol{V}$ since $\boldsymbol{U}$ is unavailable for observations in the target populations. Naively ignoring $\boldsymbol{U}$ throughout training in the source population can yield intervals with severe miscoverage when $\boldsymbol{U}$ includes important confounders between Y and A. In turn, our methods leverage $\boldsymbol{U}$ in the source population, while allowing for $\boldsymbol{U}$ to be unmeasured in the target population.

Multiple prior works have considered settings where $\mathcal{A} = \{0, 1\}$ and interest lies in constructing intervals for individual treatment effects (ITEs) $Y(1) - Y(0)$ (Lei and Candès, 2021; Yang et al., 2024; Alaa et al., 2023). We focus on constructing intervals for counterfactual outcomes over ITEs for numerous reasons. While ITEs are useful estimands in many decision-support settings, in many settings ITEs are less informative for decision-makers. For instance, when costs are associated with different treatment levels, one may prefer a less costly treatment so long as their predicted outcome is predicted to exceed a lower bound, regardless of whether the more expensive treatment would yield a greater response. Further, in settings with many treatment levels, the utility of numerous pairwise ITE intervals is less apparent. In Appendix E we demonstrate how our method can be adjusted to produce intervals for ITEs.

2.1. Assumptions

The construction of valid prediction intervals for counterfactual outcomes under runtime confounding requires a set of standard causal inference assumptions, as well as assumptions commonly invoked in causal data fusion problems (Degtiar and Rose, 2023). We begin with assumptions necessary for forming valid prediction intervals within the source population.

A1 (Positivity) $0 < \mathbb{P}(A = a|\boldsymbol{X} = \boldsymbol{x}) < 1$ *for all* $\boldsymbol{x}$ *with positive support*

A2 (Consistency) $Y = \sum_{a \in \mathcal{A}} \mathbb{I}(A = a) \cdot Y(a)$

A3 (Unconfoundedness) $Y(a) \perp\!\!\!\perp A|\boldsymbol{X}, S = 1$

Assumptions 1-3 are standard assumptions in the causal inference literature (Rosenbaum and Rubin, 1983; Rubin, 2005). While the above assumptions are sufficient for the construction of valid prediction intervals in the source data, we require additional assumptions to construct prediction intervals over the target population.

A4 (Source exchangeability) $Y(a) \perp\!\!\!\perp S|\boldsymbol{V}$

A5 (Source positivity) $0 < \mathbb{P}(S = 1|\boldsymbol{V} = \boldsymbol{v}) < 1$ *for all* $\boldsymbol{v}$ *with positive support*

Assumptions 4 and 5 are standard assumptions in the data fusion literature (Bareinboim and Pearl, 2016; Degtiar and Rose, 2023), where interest typically fixates on using source data to estimate average treatment effects for a separate target population. Assumption 4 implies all systematic differences in $Y(a)$ across the source and target population are explained by $\boldsymbol{V}$, while Assumption 5 ensures overlap of the distribution of $\boldsymbol{V}$ between the target and source populations. Collectively, the independence Assumptions 3 and 4 imply the distribution shift between observational units in the source and target populations arises due to two sources of covariate shift: (i) covariate shift in $\boldsymbol{X}$ across treatment levels within the source population, and (ii) covariate shift in $\boldsymbol{V}$ across the source and target populations. Note that by naively discarding all of $\boldsymbol{U}$ within the source population and ignoring the possibility of runtime confounding, researchers would implicitly require a more stringent variant of Assumption 3 that instead requires $Y(a) \perp\!\!\!\perp A|\boldsymbol{V}, S = 1$. Our set of Assumptions relaxes this requirement, allowing for the possibility that $\boldsymbol{U}$ is a confounder of Y and A so long as all systematic differences in counterfactual outcomes $Y(a)$ across the target and source populations are explained by the always-observed covariates $\boldsymbol{V}$.

Figure 1: Two possible directed acyclic graphs consistent with Assumptions 3 and 4. Runtime confounding is induced when $\boldsymbol{U}$ is unobserved when $S = 0$.

We note that in settings where data are collected by a single site—with an initial batch $(S = 1)$ that includes measurements on $\boldsymbol{X}$ used for training, and a second batch $(S = 0)$ only containing measurements on $\boldsymbol{V}$—Assumption 4 will often be plausible, since batch membership will commonly arise from data collection logistics, rather than factors that systematically influence the outcome of interest. This batch collection setting is exclusively considered in Coston et al. (2020), who implicitly invoke Assumption 4. We instead adopt a more general framing that treats the source and target datasets as arising from distinct sites, so as to cover broader data collection regimes. We note that the single-site batch setting considered by (Coston et al., 2020) can be viewed as a special case of this formulation, where Assumption 4 makes explicit what sources of distribution shift are permitted across batches. Like Assumption 3, Assumption 4 should be informed by subject-matter expertise in these broader settings, and we discuss avenues for sensitivity analyses in Section 6. Further, in Appendix F we show that Assumption 4 is *implied* by (i) a weaker conditional independence $Y(a) \perp\!\!\!\perp S|\boldsymbol{X}$, and (ii) a covariate shift condition on $\boldsymbol{U}$ across study populations $\boldsymbol{U} \perp\!\!\!\perp S|\boldsymbol{V}$. We additionally discuss examples of settings in which we expect Assumption 4 to hold or be violated. Figure 1 presents example causal directed acyclic graphs consistent with Assumptions 3 and 4.

3. Conformal Prediction

Conformal prediction (Vovk et al., 2005) is a general statistical framework enabling the construction of valid prediction intervals under minimal distributional assumptions. The framework's generality

has spurred a growing field of research into implementations that account for common challenges arising in prediction tasks. We provide a brief overview of the framework here, and recommend Fontana et al. (2023) and Angelopoulos et al. (2024) for extensive reviews of the many active research areas in this field.

Briefly considering a single-source prediction problem with covariates $\boldsymbol{X}$ and outcome Y for simplicity, an object central to conformal prediction is the *conformity score* $R(Y, \boldsymbol{X})$. At a high-level, $R(Y, \boldsymbol{X})$ is defined so that extreme values imply a lack of agreement, or conformity, between the actual outcome Y and predicted value based on $\boldsymbol{X}$. Conversely, smaller magnitude values imply higher conformity. Numerous conformity scores are used in practice, though two popular choices are the absolute residual $R(Y, \boldsymbol{X}) = |Y - \hat{\mathbb{E}}(Y|\boldsymbol{X})|$ and the quantile regression score $R(Y, \boldsymbol{X}) = \max\{Y - \hat{Q}_{\alpha/2}(\boldsymbol{X}), \hat{Q}_{1-\alpha/2}(\boldsymbol{X}) - Y\}$ (Romano et al., 2019), where $Q_\alpha(\boldsymbol{X})$ is the α quantile of the conditional distribution $Y|\boldsymbol{X}$.

For a fixed choice of conformity score, let $r_{1-\alpha}$ denote its theoretical $1-\alpha$ quantile. Conformal prediction is driven by the observation that for intervals of the form $\hat{C}(\boldsymbol{X}) = \{y : R(y, \boldsymbol{X}) \leq r_{1-\alpha}\}$, by construction $\mathbb{P}(Y \in \hat{C}(\boldsymbol{X})) = 1 - \alpha$. In turn, conformal prediction methods fixate interest on estimation of the unknown quantity $r_{1-\alpha}$, with many methods performing adjustments to improve finite-sample performance or, in certain circumstances, provide distribution-free finite sample guarantees that ensure a coverage lower bound of $1 - \alpha$ (Angelopoulos and Bates, 2023). Such finite-sample guarantees are generally unattainable when there is covariate shift between the source and target dataset whose form must be estimated (Barber et al., 2023). To reconcile the impossibility of finite-sample guarantees, our proposed methods leverage tools from semiparametric theory to enhance their finite sample behavior.

3.1. Constructing Conformity Scores Under Runtime Confounding

While our proposed methods are valid for any choice of conformity score, we briefly present recommendations for constructing conformity scores under runtime confounding. In such settings, conformity scores take the form $R_a(Y(a), \boldsymbol{V})$, reflecting the partial availability of covariates in the target population and fact that scores will be indexed by counterfactual outcomes occurring under different treatment levels $a \in \mathcal{A}$. Notice $R_a(Y(a), \boldsymbol{V})$ will only be observable for source units with treatment level $A = a$, for whom the consistency Assumption 2 implies $R_a(Y(a), \boldsymbol{V}) = R_a(Y, \boldsymbol{V})$. In turn, we interchangeably use both notations as appropriate. For brevity, we consider analogues of the absolute residual and quantile conformity scores.

Under Assumptions 1-5, Coston et al. (2020), who focus on counterfactual prediction of $Y(a)$ under runtime confounding, leverage the observation that $\mathbb{E}[Y(a)|\boldsymbol{V}, S = 0] = \mathbb{E}[\mathbb{E}(Y|A = a, \boldsymbol{X}, S = 1)|\boldsymbol{V}, S = 1]$. This observation implies one can construct point predictors of $Y(a)$ by (i) estimating $\mu_a(\boldsymbol{X}) = \mathbb{E}(Y|A = a, \boldsymbol{X}, S = 1)$ within the source population, and (ii) further regressing $\hat{\mu}_a(\boldsymbol{X})$ on $\boldsymbol{V}$ within the source population, obtaining predictions $\hat{\eta}_a(\boldsymbol{V})$. Such a procedure implies the conformity score $R_a(Y(a), \boldsymbol{V}) = |Y(a) - \hat{\eta}_a(\boldsymbol{V})|$, computable for source population units receiving treatment level $A = a$. Coston et al. (2020) additionally propose a doubly-robust procedure that enables faster convergence, but they do not focus on the construction of prediction intervals. In either case, our proposed methods in Section 4 provide a valid procedure to quantify the uncertainty around the resulting predictions.

We next propose a method for constructing quantile conformity scores, which relies on estimation of the quantile function of the conditional distribution $Y(a)|\boldsymbol{V}, S = 0$. Let $Q_{a,\alpha}(\boldsymbol{X})$ and

$Q_{a,\alpha}(\boldsymbol{V})$ denote the $1-\alpha$ quantiles of $Y(a)|\boldsymbol{X}, S=0$ and $Y(a)|\boldsymbol{V}, S=0$, respectively. Estimation of $Q_{a,\alpha}(\boldsymbol{V})$ requires additional care, since $\mathbb{E}[Q_{a,\alpha}(\boldsymbol{X})|\boldsymbol{V}, S=1] \neq Q_{a,\alpha}(\boldsymbol{V})$ due to the nonlinearity of quantile functions. The result below provides a valid loss function for estimation of $Q_{a,\alpha}(\boldsymbol{V})$ in the presence of runtime confounding.

Proposition 1 *Suppose $Q_{a,\alpha}(\boldsymbol{v})$ satisfies $\mathbb{P}(Y(a) \leq Q_{a,\alpha}(\boldsymbol{v})|\boldsymbol{V}=\boldsymbol{v}, S=0) = 1-\alpha$ for all $\boldsymbol{v}$ with positive support. Then, under Assumptions 1-5, $Q_{a,\alpha}(\boldsymbol{V})$ additionally satisfies*

$$Q_{a,\alpha}(\boldsymbol{V}) = \underset{\tilde{Q}_{a,\alpha}}{\arg\min}\, \mathbb{E}\left[w_a(\boldsymbol{O})\rho_\alpha(Y - \tilde{Q}_{a,\alpha}(\boldsymbol{V}))\right], \tag{1}$$

where $\rho_\alpha(x) = \alpha|x|\mathbb{I}(x \geq 0) + (1-\alpha)|x|\mathbb{I}(x < 0)$ is the pinball loss function and

$$w_a(\boldsymbol{O}) := \frac{\mathbb{I}(A=a)S(1-\kappa(\boldsymbol{V}))}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}.$$

where $g_a(\boldsymbol{X}) := \mathbb{P}(A=a|\boldsymbol{X}, S=1)$ and $\kappa(\boldsymbol{V}) := \mathbb{P}(S=1|\boldsymbol{V})$. Proposition 1 suggests one can consistently estimate the conditional quantile function $Q_{a,\alpha}(\boldsymbol{V})$ through minimization of the empirical analogue of the weighted pinball loss function (Koenker and Bassett Jr, 1978) appearing in Proposition 1. Notably, (1) represents a weighted quantile regression among source units with treatment level $A=a$, enabling the use of existing software packages which support weights or simple augmentations to existing estimation procedures. Intuitively, the weight $w_a(\boldsymbol{O})$ can be interpreted as a product of two distinct adjustment factors accounting for two sources of covariate shift. The first, $1/g_a(\boldsymbol{X})$, is an inverse probability of treatment weight that adjusts for covariate shift in $\boldsymbol{X}$ across treatment levels within the source population. The second, $(1-\kappa(\boldsymbol{V}))/\kappa(\boldsymbol{V})$, is an inverse odds weight that re-weights the source population to resemble the target population with respect to $\boldsymbol{V}$, accounting for covariate shift in $\boldsymbol{V}$ across these two populations. Relative to quantile scores based on unweighted quantile regression, we expect intervals based on scores formed from our proposed weighted quantile loss function to exhibit improved finite-sample performance, since the weights $w_a(\boldsymbol{O})$ enable consistent estimation of $Q_{a,\alpha}(\boldsymbol{V})$.

4. Methods

Given a fixed conformity score $R_a(Y, \boldsymbol{V})$, we aim to construct intervals of the form $\hat{C}_a(\boldsymbol{V}) = \{y : R_a(y, \boldsymbol{V}) \leq \hat{r}_{a,\alpha}\}$, where $\hat{r}_{a,\alpha}$ is an estimate of the $1-\alpha$ quantile of scores R_a in the target distribution which satisfies

$$\mathbb{P}(R_a(Y(a), \boldsymbol{V}) \leq \hat{r}_{a,\alpha}|S=0) = 1-\alpha. \tag{2}$$

(2) implies that construction of valid prediction intervals for $Y(a)$ in the target population requires accurate estimation of $r_{a,\alpha}$. In this Section, we present a roadmap for constructing efficient estimators of $r_{a,\alpha}$.

4.1. Identification

Although conformity scores cannot be directly observed within the target population, Assumptions 1-5 ensure $r_{a,\alpha}$ can be expressed in terms of scores formed in the source population. We present two novel identification functionals which enable estimation of $r_{a,\alpha}$ in Theorem 2 below.

Theorem 2 *Let $R_a(Y, \boldsymbol{V})$ be a generic conformity score for $Y(a)$, and suppose $r_{a,\alpha}$ satisfies*

$$\mathbb{P}(R_a(Y(a), \boldsymbol{V}) \leq r_{a,\alpha} | S = 0) = 1 - \alpha.$$

Under Assumptions 1-5, $r_{a,\alpha}$ additionally satisfies

$$\mathbb{E}[m_a(r_{a,\alpha}, \boldsymbol{V}) \mid S = 0] = 1 - \alpha \tag{3}$$
$$\mathbb{E}\left[w_a(\boldsymbol{O})\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha})\right] = 1 - \alpha. \tag{4}$$

where we define $q_a(r, \boldsymbol{X}) := \mathbb{P}(R_a(Y, \boldsymbol{V}) \leq r | \boldsymbol{X}, A = a, S = 1)$ and $m_a(r, \boldsymbol{V}) := \mathbb{E}[q_a(r, \boldsymbol{X}) | \boldsymbol{V}, S = 1]$. Equations (3) and (4) are closely related to functionals arising in the transportability literature (Zeng et al., 2025), and can be viewed as extensions of identifying functionals from the conformal prediction under covariate shift literature (Tibshirani et al., 2019; Yang et al., 2024). Intuitively, the above expressions both address two separate sources of covariate shift: between levels of treatment A among source population units, and between members of the source and target populations. Critically, (3) and (4) suggest regression- and weighting-based means for constructing valid prediction intervals of $Y(a)$ in the target population.

4.2. Plug-in Estimation

Given the identification expression (3), a natural approach to constructing prediction intervals for $Y(a)$ is to form a plug-in estimate of $r_{a,\alpha}$ by choosing $\hat{r}_{a,\alpha}$ to solve the following estimating equation in r

$$\frac{1}{n_0} \sum_{i:S_i=0} \hat{m}_a(r, \boldsymbol{V}_i) - (1 - \alpha) = 0, \tag{5}$$

noting solving (5) requires repeatedly fitting $\hat{q}_a(r, \boldsymbol{X})$ and $\hat{m}_a(r, \boldsymbol{V})$ with pre-specified learners for potentially many values of r. For any r, $m_a(r, \boldsymbol{V})$ can be estimated by first fitting $\hat{q}_a(r, \boldsymbol{X})$ among units with treatment $A = a$ in the source population, regressing the predicted values on $\boldsymbol{V}$ among source units. Letting $\mathbb{P}_n\{f(\boldsymbol{O})\} := \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{O}_i)$ for generic f, one can additionally obtain a weighting-based estimator of $r_{a,\alpha}$ through (4) by solving the estimating equation

$$\mathbb{P}_n\{\hat{w}_a(\boldsymbol{O})\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha})\} - (1 - \alpha) = 0, \tag{6}$$

which circumvents the computational challenge faced by (5) since $w_a(\boldsymbol{O})$ does not depend on r. While the above plug-in estimators are consistent for $r_{a,\alpha}$ under correct specification of all relevant nuisance models, the *rate* at which these estimators converge to $r_{a,\alpha}$ will be dictated by the convergence rates of their corresponding nuisance function estimators. These rates can be particularly slow when one chooses to fit all nuisance models with flexible learners, in order to minimize the risk of model misspecification. Following Zeng et al. (2025), Gao et al. (2025) and Liu et al. (2024), we propose the use of multiply-robust estimators of $r_{a,\alpha}$ that enable faster convergence rates in broader estimation settings. We turn our attention to the construction of these estimators in the remainder of this Section.

4.3. Efficiency Theory

In this section, we present the efficient influence curve (EIC) for $r_{a,\alpha}$. EICs are crucial ingredients for constructing estimators whose convergence rate is dictated by the *product* of nuisance function

Algorithm 1 Debiased machine learning split conformal prediction for runtime confounding

Input: Pooled target and source population data $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i, S_i)$, desired coverage probability $1-\alpha$, conformity score measure $R_a(Y, \boldsymbol{V})$

Output: A prediction interval function $\hat{C}_a(\boldsymbol{V})$

1. Randomly split the data into a training and calibration set: $\mathcal{D}_{\text{train}} = \{O_i = (Y_i, A_i, \boldsymbol{V}_i, \boldsymbol{U}_i, S_i),\ i \in \mathcal{I}_{\text{train}}\}$, $\mathcal{D}_{\text{cal}} = \{\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{V}_i, \boldsymbol{U}_i, S_i),\ i \in \mathcal{I}_{\text{cal}}\}$. Further split $\mathcal{D}_{\text{train}}$ into equally-sized subsets $\mathcal{D}_{\text{train},1}$ and $\mathcal{D}_{\text{train},2}$

2. Using all of $\mathcal{D}_{\text{train}}$, fit the nuisance functions $\hat{g}_a$ and $\hat{\kappa}$ and perform counterfactual prediction of $Y(a)$ to construct conformity scores $R_a(Y, \boldsymbol{V})$. Using $\mathcal{D}_{\text{train},1}$, obtain an initial estimate of $r_{a,\alpha}$, termed $\hat{r}^{\text{init}}_{a,\alpha}$, through a generic estimation algorithm such as weighted conformal prediction (Tibshirani et al., 2019)

3. Using $\hat{r}^{\text{init}}_{a,\alpha}$ and the second training subset $\mathcal{D}_{\text{train},2}$, obtain estimates $\hat{q}_a(\hat{r}^{\text{init}}_{a,\alpha}, \boldsymbol{X})$ and $\hat{m}_a(\hat{r}^{\text{init}}_{a,\alpha}, \boldsymbol{V})$

4. Choose $\hat{r}_{a,\alpha}$ to solve the estimating equation $\sum_{i \in \mathcal{D}_{\text{cal}}} \chi_a(r_{a,\alpha}, \boldsymbol{O}_i\,; \hat{\eta}_a(\hat{r}^{\text{init}}_{a,\alpha})) = 0$ among units in $\mathcal{D}_{\text{cal}}$

5. Use the resulting estimate $\hat{r}_{a,\alpha}$ to construct conformal intervals for participants in the target population of the form $\hat{C}_a(\boldsymbol{V}) = \{y : R_a(y, \boldsymbol{V}) \leq \hat{r}_{a,\alpha}\}$

convergence rates, often allowing for significantly faster estimation of statistical functionals relative to plug-in estimation (Kennedy 2024) while providing partial protection against model misspecification (Chernozhukov et al., 2018). The EIC for $r_{a,\alpha}$ in a fully nonparametric statistical model is presented below, with details on its derivation provided in Appendix A.

Theorem 3 *Let $\eta_a(r) := (q_a(r), m_a(r), g_a, \kappa)$ and suppose $\boldsymbol{O} \sim \mathbb{P}$. Under Assumptions 1-5, the efficient influence curve for $r_{a,\alpha}$ in a nonparametric model for the observed data distribution $\mathbb{P}$ is proportional to*

$$\begin{aligned}
&\chi_a(r_{a,\alpha}, \boldsymbol{O}\,; \eta_a(r_{a,\alpha})) := \\
&(1-S)(m_a(r_{a,\alpha}, \boldsymbol{V}) - (1-\alpha)) + \frac{S(1-\kappa(\boldsymbol{V}))}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}, \boldsymbol{X}) - m_a(r_{a,\alpha}, \boldsymbol{V})\} \\
&+ w_a(\boldsymbol{O})\{\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha}) - q_a(r_{a,\alpha}, \boldsymbol{X})\},
\end{aligned} \tag{7}$$

where $\mathbb{E}[\chi_a(r_{a,\alpha}, \boldsymbol{O}\,; \eta_a(r_{a,\alpha}))] = 0$.

Since the true EIC is mean-zero, we omit the proportionality constant when presenting χ_a since we will ultimately leverage this moment condition to construct estimators for $r_{a,\alpha}$.

4.4. Debiased Machine Learning Estimation

The efficient influence curve from Theorem 3 can be leveraged to construct efficient estimators of $r_{a,\alpha}$. We follow the framework of *debiased machine learning*, where one chooses $\hat{r}_{a,\alpha}$ to solve an estimating equation implied by the moment condition in Theorem 3 (Kennedy, 2024).

The moment condition $\mathbb{E}[\chi_a(r_{a,\alpha}, \boldsymbol{O}; \eta_a(r_{a,\alpha}))] = 0$ from Theorem 3 suggests one can obtain a valid estimate of $r_{a,\alpha}$ by choosing $\hat{r}_{a,\alpha}$ to solve $\mathbb{P}_n\{\chi_a(\hat{r}_{a,\alpha}, \boldsymbol{O}\,; \hat{\eta}_a(\hat{r}_{a,\alpha}))\} = 0$. However, naively solving the estimating equation in this manner would require iteratively estimating the nuisance functions q_a and m_a for potentially many values of r. To avoid the computational costs associated with this approach, we follow Gao et al. (2025) and construct debiased estimators for $r_{a,\alpha}$ based on the DML framework from Kallus et al. (2024) which allows for q_a and m_a to be estimated at a single, initial estimate of $r_{a,\alpha}$, drastically reducing the computational costs that would be required from repeated estimation of these nuisance functions. This localized construction introduces a preliminary estimator $\hat{r}^{\text{init}}_{a,\alpha}$, for which $n^{-1/4}$-consistency suffices (Kallus et al., 2024). By contrast, a non-localized implementation would directly re-estimate $q_a(r, \cdot)$ and $m_a(r, \cdot)$ across candidate values of r, so no separate initial estimator is needed. Our proposed split conformal prediction approach, which yields an efficient estimator $\hat{r}_{a,\alpha}$ and corresponding prediction interval $\hat{C}_a(\boldsymbol{V})$ is outlined in Algorithm 1.

While Algorithm 1 employs split conformal prediction for computational tractability, our framework extends to full conformal prediction by solving $\mathbb{P}_n\chi_a(r, \boldsymbol{O}; \hat{\eta}_a)$ without data splitting, as in Yang et al. (2024). This avoids efficiency losses from sample splitting but requires Donsker-type regularity conditions on the nuisance function classes along with the aforementioned increased computational cost.

4.5. Coverage Properties

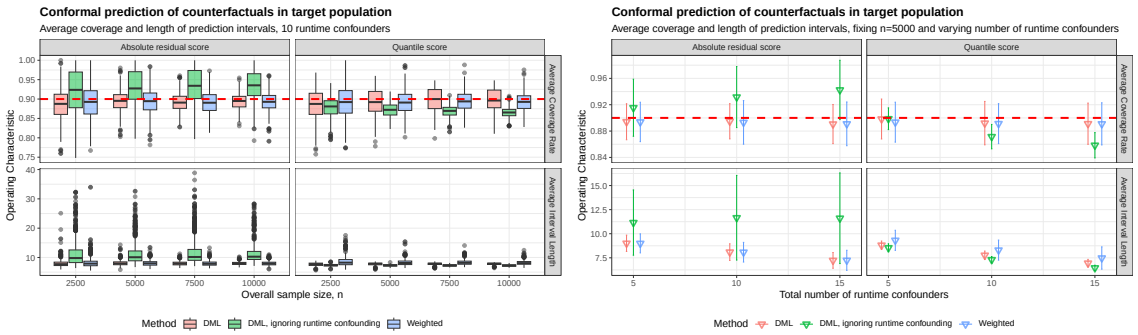
Given a means to construct prediction intervals, we turn our attention to the asymptotic coverage properties of our proposed methods. Since the formation of prediction intervals in our proposed setting requires estimation of the quantity $r_{a,\alpha}$ in (2), one would expect accurate estimation of $r_{a,\alpha}$ should ensure valid coverage of $Y(a)$. Theorem 4 provides a formal characterization of this notion.

Theorem 4 *Suppose that $\hat{\kappa}(\boldsymbol{V})$, $\hat{g}_a(\boldsymbol{X})$ and $\mathbb{P}(S = 1|\boldsymbol{X})$ are all bounded within $(\varepsilon, 1-\varepsilon)$ for some $\varepsilon > 0$. If $\hat{C}_a(\boldsymbol{V})$ is constructed according to Algorithm 1, then*

$$\mathbb{P}(Y(a) \in \hat{C}_a(\boldsymbol{V})|S = 0) = 1 - \alpha + O_{\mathbb{P}}(1/\sqrt{n} \;+\; R_n), \text{ where}$$

$$R_n = \sup_r ||\hat{q}_a(r, \cdot) - q_a(r, \cdot)|| \cdot ||\hat{g}_a - g_a|| + \sup_r ||\hat{m}_a(r, \cdot) - m_a(r, \cdot)|| \cdot ||\hat{\kappa} - \kappa||.$$

The structure of the remainder term R_n implies the coverage error shrinks quickly as long as either the estimated outcome-related models (q_a, m_a) or the propensity score models (g_a, κ) converge sufficiently fast. For example, if all four nuisance functions are estimated using flexible learners with modest convergence rates of $n^{-1/4}$, the product of errors will be of order $n^{-1/2}$. This implies the coverage gap shrinks at the parametric $n^{-1/2}$ rate—the fastest possible rate when $\eta_a(r)$ must be estimated (Kennedy, 2024) and a dramatic improvement over plug-in estimators, whose convergence would be limited to the slower $n^{-1/4}$ rate. Due to the protections against misspecification and slow convergence rates, estimators arising from the DML framework are frequently referred to as *doubly-* or *multiply-robust* (Kennedy, 2024). Beyond the rate conditions on R_n, Theorem 4 implies a model-multiple robustness property: consistency of $\hat{r}_{a,\alpha}$ is guaranteed whenever at least one nuisance function in each of the pairs (m_a, κ) and (q_a, g_a) is consistently estimated, irrespective of the convergence behavior of the remaining components. Such improvements are crucial in runtime confounding settings, where numerous nuisance functions need to be modeled.



(*a*) (*b*)

Figure 2: Experiment results, varying n and fixing number of runtime confounders at 10.

5. Experiments

5.1. Simulated Data

To assess the performance of our proposed methods, we conducted a simulation study extending the setup considered in Coston et al. (2020), who focused on efficient point prediction of $Y(a)$ under the same runtime confounding setting we study. We generate data according to the runtime confounding data structure implied by Figure 1 and Table 1, letting $\mathcal{A} = \{0, 1\}$ for simplicity and recalling our proposed methods can accommodate categorical $\mathcal{A}$. Additionally, we vary the overall sample size n and number of unmeasured runtime confounders (5, 10, 15), corresponding to cases of mild, moderate and severe runtime confounding. Throughout, we generate S so that $\mathbb{P}(S = 1) = 0.9$, implying for each sample size considered 90% of observations are from the source population. We extend Coston et al. (2020) by generating S as a function of $\boldsymbol{V}$ to ensure covariate shift across the source and target populations.

Full details on the data-generating mechanism, which produces data adhering to the structure in Table 1, can be found in Appendix B. Replication code can be found at `https://github.com/keithbarnatchez/conformal-runtime`. Additional experiments which vary the relative size of the source population are included in Appendix C. Across the simulation settings we explore, we construct 90% prediction intervals for both $Y(1)$ and $Y(0)$ in the target data based on both absolute residual and quantile conformity scores. We compared our proposed DML procedure to (i) the weighted method implied by equation (4) and (ii) a DML estimator based on the approach from Yang et al. (2024) which ignores runtime confounding by effectively forcing $\boldsymbol{V} = \boldsymbol{X}$. These two approaches serve as natural comparators, since (i) is the standard approach to addressing distribution shift in conformal prediction problems, and (ii) allows us to investigate the consequences of ignoring runtime confounding while employing an analogous estimation procedure. While not the main contribution of our work, we emphasize that the weighted approach is based on our proposed weights $\hat{w}_a(\boldsymbol{O})$, and in turn can be viewed as an additional approach for addressing runtime confounding that we provide.

Results: Figure 2 displays the results of our experiments. For brevity, we report the empirical coverage rates and interval lengths for $Y(1)$ and $Y(0)$ pooled together, and provide results separately for $Y(1)$ and $Y(0)$ in Appendix B. We first focus on panel (a), which considers the moderate runtime confounding setting while varying n. We see that naively ignoring runtime confounding produces intervals which miscover $Y(1)$ and $Y(0)$ at all sample sizes considered. Notably, as the

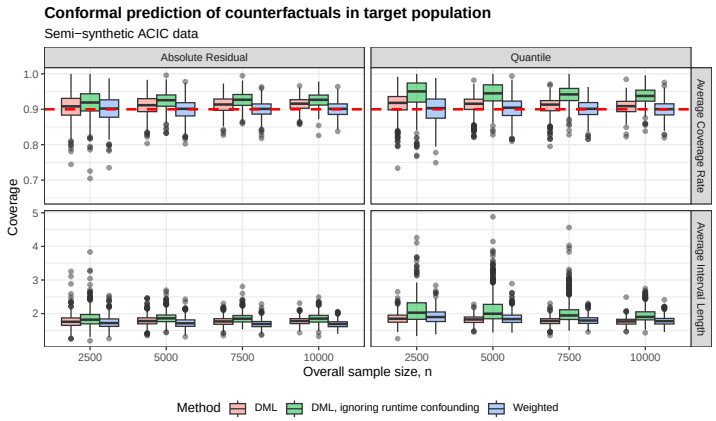


Figure 3: Performance of proposed methods on semi-synthetic ACIC data.

overall sample size grows our proposed method rapidly concentrates around the desired 90% coverage rate, consistent with the results of Theorem 4. Focusing on panel (b), we see that as runtime confounding becomes more severe, the coverage bias of the DML procedure which ignores runtime confounding worsens, with opposite magnitudes of bias across the two considered conformity scores, highlighting that the coverage bias induced by ignoring runtime confounding can vary systematically and demonstrating the need to adjust for runtime confounding in practice. Across all levels of runtime confounding and both conformity scores considered, intervals based on both our DML procedure and weighted conformal with our proposed weights $\hat{w}_a(\boldsymbol{O})$ both consistently attain the desired coverage rate of 90%. Notably, our DML procedure tends to produce intervals that are as or more narrow than the weighted conformal procedure and concentrates rapidly around the nominal 90% rate as n grows, highlighting the efficiency of our proposed approach.

5.2. Semi-Synthetic Data

We examine the performance of our proposed methods on semi-synthetic data from the 2018 Atlantic Causal Inference Conference (ACIC) challenge (Carvalho et al., 2019), which is based on the National Study of Learning Minds (NSLM) trial (Yeager et al., 2019) and has been used in previous studies of conformal inference for counterfactual outcomes (Lei and Candès, 2021). To ensure access to ground-truth counterfactual outcomes, we generate 1,000 synthetic datasets from the ACIC NSLM database following the same approach outlined in Lei and Candès (2021), who used this dataset to evaluate weighted conformal prediction methods for individual treatment effects. We enforce runtime confounding by simulating a source population indicator dependent on a subset of covariates in the ACIC NSLM dataset, and assume the analyst only has access to this subset of covariates for the target population. Analogous to Section 5.1 we fix $\mathbb{P}(S = 1) = 0.9$, and repeat this exercise for different overall sample sizes, and vary variables included in $\boldsymbol{V}$ and $\boldsymbol{U}$ to examine increasingly severe cases of runtime confounding that we term mild, moderate and severe. We generate S as a function of $\boldsymbol{V}$ that enforces covariate shift between the target and source populations. Full details on the data generation procedure are provided in Appendix B.

Results: Figure 3 displays the results of our exercise for the moderate runtime confounding scenario. Results for the mild and severe scenarios are qualitatively similar and reported in Appendix D. Similar to our numerical experiments in Section 5.1, we see that our proposed DML procedure and the weighted conformal approach based on our derived weights $w_a(\boldsymbol{O})$ both attain approximately

valid coverage. Naively applying Yang et al. (2024) and ignoring runtime confounding continues to lead to miscoverage that is most pronounced when using quantile scores. Along with possessing the largest coverage bias, the naive approach consistently produces the widest prediction intervals, further demonstrating the consequences that can arise from ignoring runtime confounding. Notably, the weighted procedure based on our derived weights $\hat{w}_a(\boldsymbol{O})$ performs well over all sample sizes considered. We suspect the slight coverage bias for our proposed DML method arises from relative complexity in the underlying conditional score functions q_a and m_a, recalling the weighted procedure does not require estimates of these functions. Consistent with Theorem 4, the proposed DML procedure attains approximately valid coverage throughout, since accurate estimation of the nuisance functions comprising $\hat{w}_a(\boldsymbol{O})$ partially protects against inaccurate estimation of q_a and m_a.

6. Discussion

We developed computationally and statistically efficient methods to construct prediction intervals for counterfactual outcomes under runtime confounding, a setting that involves both treatment-outcome confounding and covariate shift between source and target populations. Our approach uses a multiply robust debiased machine learning estimator of the required conformity score quantile, enabling the resulting prediction intervals to achieve desired coverage rates under modest nuisance learning requirements. Our theoretical results show this coverage is achieved more rapidly as a function of n than with standard plug-in methods, and numerical experiments identify numerous scenarios where our method displays superior or comparable performance to standard approaches. Additionally, we provided valid loss functions for performing counterfactual quantile regression in runtime confounding settings, and a weighted conformal prediction method that effectively addressed runtime confounding bias throughout our numerical experiments. Both the proposed DML method and the weighted procedure consistently outperform a state-of-the-art DML procedure that ignores runtime confounding in our numerical experiments, highlighting the need to address runtime confounding in practice.

A limitation of our approach is that the validity of our method relies on the causal and transportability Assumptions 1-5, which are untestable. The procedure may fail if, for example, there is an unmeasured confounder of the treatment-outcome relationship or if the source and target populations differ in ways not captured by the observed covariates $\boldsymbol{V}$. Future work could extend our framework to several important areas. One direction is developing formal sensitivity analyses to quantify how prediction intervals and coverage rates are affected by violations of the core independence Assumptions 3-4. Extending the sensitivity analysis framework from Zeng et al. (2025), who focused on ATE estimation under runtime confounding, could serve as a promising avenue forward. Additionally, our framework could be extended to support continuous treatments in runtime confounding settings (Schröder et al., 2025) and survival outcomes (Candes et al., 2023).

References

Ahmed M Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. *Advances in neural information processing systems*, 36: 47682–47703, 2023.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. *Advances in Neural Information Processing Systems*, 35:37184–37198, 2022.

Alexander Bleier, Avi Goldfarb, and Catherine Tucker. Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing*, 37(3):466–480, 2020.

Christopher B Boyer, Issa J Dahabreh, and Jon A Steingrimsson. Estimating and evaluating counterfactual prediction models. *Statistics in Medicine*, 44(23-24):e70287, 2025.

Emmanuel Candes, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5 (2):21–35, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, pages C1–C68, 2018.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on fairness, accountability and transparency*, pages 134–148. PMLR, 2018.

Gary S Collins and Paula Dhiman. Prediction models should contain predictors known at the moment of intended use. *Aging Clinical and Experimental Research*, 35(12):3243–3244, 2023.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual predictions under runtime confounding. *Advances in neural information processing systems*, 33:4150–4162, 2020.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.

Mieke Deschepper, Chloë De Smedt, and Kirsten Colpaert. A literature-based approach to predict continuous hospital length of stay in adult acute care patients using admission variables: A single university center experience. *International Journal of Medical Informatics*, 193:105678, 2025.

Unai Fischer-Abaigar, Christoph Kern, Noam Barda, and Frauke Kreuter. Bridging the gap: Towards an expanded toolkit for ai-driven decision-making in the public sector. *Government Information Quarterly*, 41(4):101976, 2024.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Chenyin Gao, Peter B Gilbert, and Larry Han. Bridging fairness and efficiency in conformal inference: A surrogate-assisted group-clustered approach. In *Forty-second International Conference on Machine Learning*, 2025.

Ellen Graham, Marco Carone, and Andrea Rotnitzky. Towards a unified theory for semiparametric data fusion with individual-level data. *arXiv preprint arXiv:2409.09973*, 2024.

Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36:70453–70482, 2023.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2025.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25 (16):1–59, 2024.

Edward H Kennedy. Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics*, 16(1), 2020.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

Ruth H Keogh and Nan Van Geloven. Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. *Epidemiology*, 35(3):329–339, 2024.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

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

Roderick J Little and Donald B Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, 21(1): 121–145, 2000.

Yi Liu, Alexander Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *International Conference on Machine Learning*, pages 31344–31382. PMLR, 2024.

Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine*, pages 795–840. Springer, 2021.

Ji Won Park and Kyunghyun Cho. Semiparametric conformal prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 3880–3888. PMLR, 2025.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

Maresa Schröder, Dennis Frauen, Jonas Schweisthal, Konstantin Hess, Valentyn Melnychuk, and Stefan Feuerriegel. Conformal prediction for causal effects of continuous treatments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Cathy Shyr, Boyu Ren, Prasad Patil, and Giovanni Parmigiani. Multi-study r-learner for estimating heterogeneous treatment effects across studies using statistical machine learning. *Biostatistics*, 26(1):kxaf040, 2025.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

AW van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Sarah C Voter, Issa J Dahabreh, Christopher B Boyer, Habib Rahbar, Despina Kontos, and Jon A Steingrimsson. Counterfactual prediction from machine learning models: transportability and joint analysis for model development and evaluation using multi-source data. *Diagnostic and Prognostic Research*, 9(1):22, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.

David S Yeager, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S Hulleman, Cintia P Hinojosa, et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774): 364–369, 2019.

Zhenghao Zeng, Edward H. Kennedy, Lisa M. Bodnar, and Ashley I. Naimi. Efficient generalization and transportation. *Statistical Science*, 40(3):495–514, August 2025.

Appendix A. Proofs

A.1. Lemmas

Lemma 5 *Let $h(Y(a), \boldsymbol{V})$ be a generic square integrable function. Under Assumptions 1-5 we have that*

$$\mathbb{E}[h(Y(a), \boldsymbol{V})|S=0] = \mathbb{E}\left(\mathbb{I}(A=a)Sh(Y, \boldsymbol{V})\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right).$$

To prove Lemma 5—which we in turn leverage in the proofs of Proposition 1 and Theorem 2—notice

$$\begin{aligned}
\mathbb{E}\left(\mathbb{I}(A=a)Sh(Y(a), \boldsymbol{V})\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right) &= \mathbb{E}\left(\mathbb{E}[\mathbb{I}(A=a)Sh(Y(a), \boldsymbol{V})|\boldsymbol{X}]\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right) \\
&= \mathbb{E}\left(\mathbb{E}[h(Y(a), \boldsymbol{V})|\boldsymbol{X}]g_a(\boldsymbol{X})\mathbb{P}(S=1|\boldsymbol{X})\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right) \\
&= \mathbb{E}\left(\mathbb{E}[h(Y(a), \boldsymbol{V})|\boldsymbol{X}]\mathbb{P}(S=1|\boldsymbol{X})\frac{1-\kappa(\boldsymbol{V})}{\kappa(\boldsymbol{V})}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left\{Sh(Y(a), \boldsymbol{V})\cdot\frac{1-\kappa(\boldsymbol{V})}{\kappa(\boldsymbol{V})}\middle|\boldsymbol{X}\right\}\right) \\
&= \mathbb{E}\left(Sh(Y(a), \boldsymbol{V})\cdot\frac{1-\kappa(\boldsymbol{V})}{\kappa(\boldsymbol{V})}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left\{Sh(Y(a), \boldsymbol{V})\cdot\frac{1-\kappa(\boldsymbol{V})}{\kappa(\boldsymbol{V})}\middle|\boldsymbol{V}\right\}\right) \\
&= \mathbb{E}\left(h(Y(a), \boldsymbol{V})|S=0\right).
\end{aligned}$$

A.2. Proof of Proposition 1

Let $Q_{a,\alpha}(\boldsymbol{v})$ denote the $1-\alpha$ quantile of $Y(a)$ conditional on $\boldsymbol{V}=\boldsymbol{v}$, and note $Q_{a,\alpha}(\boldsymbol{v})$ satisfies (Koenker and Bassett Jr, 1978)

$$\mathbb{P}(Y(a) \leq Q_{a,\alpha}(\boldsymbol{V})|\boldsymbol{V}, S=0) = 1-\alpha.$$

Letting $\rho_\alpha(x) := \alpha|x|\mathbb{I}(x>0) + (1-\alpha)|x|\mathbb{I}(x\leq 0)$ denote the pinball loss function, notice $Q_{a,\alpha}(\boldsymbol{V})$ additionally satisfies

$$Q_{a,\alpha}(\boldsymbol{V}) = \underset{\tilde{Q}_{a,\alpha}}{\arg\min}\,\mathbb{E}[\rho_\alpha\{Y(a) - Q_{a,\alpha}(\boldsymbol{V})\}|S=0]$$

For brevity, let $h_\alpha(\boldsymbol{V}, Y(a); Q) := \rho_\alpha\{Y(a) - Q(\boldsymbol{V})\}$, and notice

$$\mathbb{E}[h_\alpha(\boldsymbol{V}, Y(a); Q)|S=0] = \mathbb{E}\left[\mathbb{I}(A=a)Sh_\alpha(\boldsymbol{V}, Y(a); Q)\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right]$$

by Lemma 5, which implies

$$Q_{a,\alpha}(\boldsymbol{V}) = \underset{\tilde{Q}_{a,\alpha}}{\arg\min}\,\mathbb{E}\left[\mathbb{I}(A=a)Sh_\alpha(\boldsymbol{V}, Y(a); \tilde{Q}_{a,\alpha})\frac{1-\kappa(\boldsymbol{V})}{g_a(\boldsymbol{X})\kappa(\boldsymbol{V})}\right],$$

proving Proposition 1 and suggesting one can estimate conditional quantiles of $Y(a)|\boldsymbol{V}$ through a simple reweighting of the conventional quantile regression loss function.

A.3. Proof of Theorem 2

We begin by proving (3). For compactness, let $R_a := R_a(Y, \boldsymbol{V})$ and notice

$$\begin{aligned}
\mathbb{P}(R_a \leq r_{a,\alpha} | S=0) &= \mathbb{E}(\mathbb{I}(R_a \leq r_{a,\alpha}) | S=0) \\
&= \mathbb{E}[\mathbb{E}(\mathbb{I}(R_a \leq r_{a,\alpha}) | \boldsymbol{X}, S=0) | S=0] \\
&= \mathbb{E}[\mathbb{E}(\mathbb{I}(R_a \leq r_{a,\alpha}) | \boldsymbol{X}, S=1) | S=0] \\
&= \mathbb{E}\{\mathbb{E}[\mathbb{E}(\mathbb{I}(R_a \leq r_{a,\alpha}) | \boldsymbol{X}, A=a, S=1) | \boldsymbol{V}, S=1] | S=0\}
\end{aligned}$$

Above, line 3 holds by Assumption 4, recalling $\boldsymbol{V} \subset \boldsymbol{X}$, and line 4 holds due to Assumptions 3 and 4. The desired result holds by recalling the definitions of $q_a(r_{a,\alpha}, \boldsymbol{X})$ and $m_a(r_{a,\alpha}, \boldsymbol{V})$.

Letting $h(Y(a), \boldsymbol{V}) := I(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha})$, Equation (4) immediately follows by Lemma 5.

A.4. Proof of Theorem 3

Suppose $\boldsymbol{O} \sim \mathbb{P}$, and let $\{\mathbb{P}_\varepsilon : \varepsilon \in [0,1)\}$ be a generic regular parametric submodel containing the true data-generating distribution at $\varepsilon = 0 : \mathbb{P}_0 = \mathbb{P}$. Recall that an *influence curve* for a pathwise differentiable functional $\Psi(\mathbb{P})$ is a function $\chi(\boldsymbol{O}, \mathbb{P})$ which satisfies

$$\frac{d}{d\varepsilon}\Psi(\mathbb{P}_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}_{\mathbb{P}}[\chi(\boldsymbol{O}, \mathbb{P}) u(\boldsymbol{O})], \tag{8}$$

for any parametric submodel, where $\mathbb{E}_{\mathbb{P}}[\chi(\boldsymbol{O}, \mathbb{P})] = 0$, $\text{Var}_{\mathbb{P}}(\chi(\boldsymbol{O}, \mathbb{P})) < \infty$ and $u(\boldsymbol{O}) = \frac{d}{d\varepsilon} \log \mathbb{P}_\varepsilon|_{\varepsilon=0}$ is the score function for the parametric submodel evaluated at $\varepsilon = 0$ (Tsiatis, 2006). For generic Q, W, let $u_{Q|W}$ be the conditional score of $Q|W$, $u_{Q,W}$ be the score function for the joint distribution of Q and W, u_W be the score function for W, and note that $u_{Q,W} = u_{Q|W} + u_W$. The tangent space is defined as the closed linear span of scores for all possible parametric submodels, and the *efficient influence curve* is the unique influence function belonging to the tangent space.

Analogous to Liu et al. (2024)—whose approach we follow—our strategy to find the efficient influence curve for $r_{a,\alpha}(\mathbb{P})$ is to begin by differentiating the identifying expression (3) induced by a generic parametric submodel $\mathbb{P}_\varepsilon$, where we will ultimately rearrange the resulting terms to arrive at an expression for $\frac{d}{d\varepsilon} r_{a,\alpha}(\mathbb{P}_\varepsilon)|_{\varepsilon=0}$. Notice we have

$$\begin{aligned}
0 &= \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}_\varepsilon}\{\,\mathbb{E}_{\mathbb{P}_\varepsilon}[\,\mathbb{E}_{\mathbb{P}_\varepsilon}(\mathbb{I}(R_a(Y,\boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P}_\varepsilon)) \mid \boldsymbol{X}, A=a, S=1) \mid \boldsymbol{V}, S=1\,] \mid S=0\,\}\Big|_{\varepsilon=0} \\
&= \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}_\varepsilon}\{\,\mathbb{E}_{\mathbb{P}}[\,\mathbb{E}_{\mathbb{P}}(\mathbb{I}(R_a(Y,\boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})) \mid \boldsymbol{X}, A=a, S=1) \mid \boldsymbol{V}, S=1\,] \mid S=0\,\}\Big|_{\varepsilon=0} \\
&+ \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}}\{\,\mathbb{E}_{\mathbb{P}_\varepsilon}[\,\mathbb{E}_{\mathbb{P}}(\mathbb{I}(R_a(Y,\boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})) \mid \boldsymbol{X}, A=a, S=1) \mid \boldsymbol{V}, S=1\,] \mid S=0\,\}\Big|_{\varepsilon=0} \\
&+ \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}}\{\,\mathbb{E}_{\mathbb{P}}[\,\mathbb{E}_{\mathbb{P}_\varepsilon}(\mathbb{I}(R_a(Y,\boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})) \mid \boldsymbol{X}, A=a, S=1) \mid \boldsymbol{V}, S=1\,] \mid S=0\,\}\Big|_{\varepsilon=0} \\
&+ \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}}\{\,\mathbb{E}_{\mathbb{P}}[\,\mathbb{E}_{\mathbb{P}}(\mathbb{I}(R_a(Y,\boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P}_\varepsilon)) \mid \boldsymbol{X}, A=a, S=1) \mid \boldsymbol{V}, S=1\,] \mid S=0\,\}\Big|_{\varepsilon=0} \\
&= \text{I} + \text{II} + \text{III} + \text{IV}
\end{aligned}$$

Focusing on the first term and recalling the definition of $m_a(r, \boldsymbol{V})$ we have

$$\begin{aligned}
\text{I} &= \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}_\varepsilon}\{m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V})|S=0\}\Big|_{\varepsilon=0} \\
&= \mathbb{E}_{\mathbb{P}}\{(m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}) - (1-\alpha))u_{\boldsymbol{V}|S=0}|S=0\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-S}{\mathbb{P}(S=0)}(m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}) - (1-\alpha))u_{\boldsymbol{V}|S}\right\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-S}{\mathbb{P}(S=0)}(m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}) - (1-\alpha))u(\boldsymbol{O})\right\}
\end{aligned}$$

Above, we are able to add in u_S since $(1-S)(m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}) - (1-\alpha))$ has mean zero given S, and in turn can then add in $u_{Y,A,\boldsymbol{U}|\boldsymbol{V},S}$, which is mean zero given $\boldsymbol{V}$ and S.

For the second term, recalling $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{U})$ and that $q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) = \mathbb{E}_{\mathbb{P}}[\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})|\boldsymbol{X}, A = a, S=1]$, notice

$$\begin{aligned}
\text{II} &= \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}_\varepsilon}[q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X})|\boldsymbol{V}]|S=0\} \\
&= \mathbb{E}_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}}[q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) - m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}))u_{\boldsymbol{U}|\boldsymbol{V},S=1}|\boldsymbol{V}, S=1]|S=0\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-S}{\mathbb{P}(S=0)}\mathbb{E}_{\mathbb{P}}[q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) - m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V}))u_{\boldsymbol{U}|\boldsymbol{V},S=1}|\boldsymbol{V}, S=1]\right\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-S}{\mathbb{P}(S=0)}\mathbb{E}_{\mathbb{P}}\left[\frac{S}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) - m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V})\}u_{\boldsymbol{U}|\boldsymbol{V},S}\right]\right\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-\kappa(\boldsymbol{V})}{\mathbb{P}(S=0)}\frac{S}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) - m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V})\}u_{\boldsymbol{U}|\boldsymbol{V},S}\right\} \\
&= \mathbb{E}_{\mathbb{P}}\left\{\frac{1-\kappa(\boldsymbol{V})}{\mathbb{P}(S=0)}\frac{S}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}) - m_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{V})\}u(\boldsymbol{O})\right\}
\end{aligned}$$

Similar to term I, on the final line we are able to add in $u_{\boldsymbol{V},S}$ since this term is mean zero given $\boldsymbol{V}$ and S. Recalling $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{U})$, we can then add in $u_{Y,A|\boldsymbol{X},S}$ following the same logic used in term I.

For the third term, following similar logic we have

$$\begin{aligned}
\text{III} &= \mathbb{E}_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{P}}[(\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})) - q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}))u_{Y|\boldsymbol{X},A=a,S=1}|\boldsymbol{X}, A=a]|\boldsymbol{V}, S=1\,|S=0\} \\
&= \mathbb{E}_{\mathbb{P}}\{\mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{P}}\left[\frac{\mathbb{I}(A=a)S}{\mathbb{P}(A=a, \boldsymbol{X}, S=1)}(\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha}(\mathbb{P})) - q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X}))u_{Y|\boldsymbol{X},A,S}\right]|\boldsymbol{V}, S=1\,|S=0\} \\
&= \mathbb{E}_{\mathbb{P}}\left[\frac{1-\kappa(\boldsymbol{V})}{\mathbb{P}(S=0)}\frac{\mathbb{I}(A=a)S}{\mathbb{P}(A=a, \boldsymbol{X}, S=1)}\{\mathbb{I}(R_a(Y, \boldsymbol{V}) - q_a(r_{a,\alpha}(\mathbb{P}), \boldsymbol{X})\}u(\boldsymbol{O})\right]
\end{aligned}$$

Finally, for the fourth term we have that

$$\begin{aligned}
\text{IV} &= \frac{d}{d\varepsilon}\mathbb{E}_{\mathbb{P}}[m_a(r_{a,\alpha}(\mathbb{P}_\varepsilon))|S=0]\Big|_{\varepsilon=0} \\
&\propto \frac{d}{d\varepsilon}r_{a,\alpha}(\mathbb{P}_\varepsilon)\Big|_{\varepsilon=0}
\end{aligned}$$

Above, we can safely ignore the proportionality constant, since influence curves are by construction mean zero and we intend to use the resulting influence curve to form estimating equation estimators of $r_{a,\alpha}(\mathbb{P})$, whose solutions are invariant to scaling. Note that if we wished to perform a one-step bias correction, we would need to incorporate this proportionality constant. Since we do not wish to perform statistical inference on $r_{a,\alpha}$, and instead simply require an efficient estimate of this quantity, there are no costs incurred by avoiding estimation of this constant.

Re-combining I, II, III and IV, and solving for $\frac{d}{d\varepsilon} r_{a,\alpha}(\mathbb{P})|_{\varepsilon=0}$, we have that

$$\begin{aligned}\frac{d}{d\varepsilon} r_{a,\alpha}(\mathbb{P})\bigg|_{\varepsilon=0} &\propto \mathbb{E}_{\mathbb{P}}\left\{\frac{1-S}{\mathbb{P}(S=0)}(m_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{V})-(1-\alpha))u(\boldsymbol{O})\right\} \\ &+ \mathbb{E}_{\mathbb{P}}\left\{\frac{1-\kappa(\boldsymbol{V})}{\mathbb{P}(S=0)}\frac{S}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{X})-m_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{V})\}u(\boldsymbol{O})\right\} \\ &+ \mathbb{E}_{\mathbb{P}}\left[\frac{1-\kappa(\boldsymbol{V})}{\mathbb{P}(S=0)}\frac{\mathbb{I}(A=a)S}{\mathbb{P}(A=a,\boldsymbol{X},S=1)}\{\mathbb{I}(R_a(Y,\boldsymbol{V})-q_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{X})\}u(\boldsymbol{O})\right]\end{aligned}$$

Noting each term above is mean zero,

$$\begin{aligned}\frac{d}{d\varepsilon} r_{a,\alpha}(\mathbb{P})\bigg|_{\varepsilon=0} &\propto \mathbb{E}_{\mathbb{P}}\Bigg[\Bigg\{(1-S)(m_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{V})-(1-\alpha))+\frac{S(1-\kappa(\boldsymbol{V}))}{\kappa(\boldsymbol{V})}\{q_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{X})-m_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{V})\} \\ &+ \frac{\mathbb{I}(A=a)S)1-\kappa(\boldsymbol{V}))}{\mathbb{P}(A=a,\boldsymbol{X},S=1)}\{\mathbb{I}(R_a(Y,\boldsymbol{V})-q_a(r_{a,\alpha}(\mathbb{P}),\boldsymbol{X})\}\Bigg\}u(\boldsymbol{O})\Bigg] \\ &= \mathbb{E}_{\mathbb{P}}[\chi_a(\boldsymbol{O},\mathbb{P};m_a,r_a,g_a,\kappa)u(\boldsymbol{O})]\end{aligned}$$

where we additionally omit the proportionality constant $\mathbb{P}(S=0)^{-1}$ initially appearing in each of the three terms above for brevity. It is then straightforward to verify that $\chi_a(\boldsymbol{O})$ is an element of the tangent space.

Recalling the definition of an EIC in (8), since $\chi_a(\boldsymbol{O},\mathbb{P};m_a,r_a,g_a,\kappa)$ is mean zero, we conclude that the efficient influence curve for $r_{a,\alpha}(\mathbb{P})$ is proportional to $\chi_a(\boldsymbol{O},\mathbb{P};m_a,r_{a,\alpha},g_a,\kappa)$.

A.5. Proof of Theorem 4

Suppose that $\hat{\eta}_a = (\hat{q}_a, \hat{m}_a, \hat{\kappa}, \hat{g})$ is obtained from a separate sample independent from $\boldsymbol{O}_i$, and assume there exists some small $\varepsilon > 0$ such that $\hat{\kappa}(\boldsymbol{V}) \in (\varepsilon, 1-\varepsilon)$, $\hat{g}_a(\boldsymbol{X}) \in (\varepsilon, 1-\varepsilon)$, and $\mathbb{P}(S=1|\boldsymbol{X}) \in (\varepsilon, 1-\varepsilon)$ almost surely.

We aim to show that

$$\mathbb{P}(Y(a) \in \hat{C}_a(\boldsymbol{V})|S=0) = 1-\alpha+O_{\mathbb{P}}(1/\sqrt{n}\ +\ R_n), \tag{9}$$

where $R_n = \sup_r ||\hat{q}_a(r,\cdot)-q_a(r,\cdot)|| \cdot ||\hat{g}_a - g_a|| + \sup_r ||\hat{m}_a(r,\cdot)-m_a(r,\cdot)|| \cdot ||\hat{\kappa}-\kappa||$. Such a construction allows for one to quantify conditions on nuisance function estimation rates such that the above coverage slack is of order $O_{\mathbb{P}}(1/\sqrt{n})$.

To achieve this, we will

1. Show $\mathbb{P}(Y(a) \in \hat{C}_a(\boldsymbol{V})|S=0) - (1-\alpha) = \mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)]/\mathbb{P}(S=0)$
2. Decompose $\mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha}\ \eta)]$ into a term whose asymptotic behavior is dominated by $\mathbb{E}(\chi_a(\hat{r}_{a,\alpha},\hat{\eta}) - \chi_a(\hat{r}_{a,\alpha},\eta))$

3. Show that for any r, the difference $\mathbb{E}(\chi_a(r,\hat{\eta}) - \chi_a(r,\eta))$ satisfies the product bias structure specified in Theorem 4

4. Take the supremum of this bias structure over all r to bound $\mathbb{E}(\chi_a(\hat{r}_{a,\alpha},\hat{\eta}) - \chi_a(\hat{r}_{a,\alpha},\eta))$

To begin, notice

$$\begin{aligned}\mathbb{P}(Y(a) \in \hat{C}_a(\boldsymbol{V})|S=0) - (1-\alpha) &= \mathbb{P}(R_a(Y(a),\boldsymbol{V}) \leq \hat{r}_{a,\alpha})|S=0) - (1-\alpha) \\ &= \mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)]/\mathbb{P}(S=0), \end{aligned} \tag{10}$$

where (10) holds since

$$\begin{aligned}\mathbb{P}(R_a(Y(a),\boldsymbol{V}) \leq r)|S=0) - (1-\alpha) &= \mathbb{E}[m_a(r,\boldsymbol{V}) - (1-\alpha)|S=0] \\ &= \mathbb{E}[\chi_a(r,O;\eta_a(r))]/\mathbb{P}(S=0),\end{aligned}$$

for any r.

Thus, demonstrating (9) amounts to showing

$$\mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)] = O_{\mathbb{P}}(1/\sqrt{n} + R_n). \tag{11}$$

We consider the following decomposition for $\mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)]$. For brevity, we omit the observational arguments and define $\mathbb{E}[\chi_a(\hat{r}_{a,\alpha},\eta)] := \mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)]$, noting

$$\begin{aligned}\mathbb{E}[\chi_a(\hat{r}_{a,\alpha},\eta)] = \mathbb{E}(\chi_a(\hat{r}_{a,\alpha},\hat{\eta}) - \chi_a(\hat{r}_{a,\alpha},\eta)) \\ - (\mathbb{P}_n - \mathbb{E})[\chi_a(\hat{r}_{a,\alpha},\hat{\eta})] \\ + \mathbb{P}_n(\chi_a(\hat{r}_{a,\alpha},\hat{\eta}))\end{aligned}$$

Above, the third term is zero by construction. The second term is $O_{\mathbb{P}}(1/\sqrt{n})$ if either (i) $\hat{\psi}_a(r_{a,\alpha},\hat{\eta})$ lies in a Donsker class (van der Vaart, 2000), or (ii) if $\hat{\eta}$ is obtained from a separate sample (Kennedy, 2020). Since we employ the cross-fitting procedure suggested by Kallus et al. (2024), condition (ii) holds regardless of whether all relevant nuisance functions fall into a Donsker class, implying the second term above is $O_{\mathbb{P}}(1/\sqrt{n})$. We note that modest assumptions on the nuisance functions $\hat{\eta}$ employed in related work (Liu et al., 2024) additionally ensure this rate of convergence without the need for cross fitting, but these assumptions are not strictly necessary given our use of cross-fitting.

We turn our focus to the first term above, $\mathbb{E}(\chi_a(\hat{r}_{a,\alpha},\hat{\eta}) - \chi_a(\hat{r}_{a,\alpha},\eta))$.

Our strategy for bounding this first term closely follows that of Zeng et al. (2025). Notice for any generic r, we have

$$\begin{aligned}&\mathbb{E}(\chi_a(r,\hat{\eta}) - \chi_a(r,\eta)) \\ &= \mathbb{E}[(1-S)(\hat{m}_a(r,\boldsymbol{V}) - m_a(r,\boldsymbol{V})] \\ &+ \mathbb{E}\left[\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}(\tilde{m}_a(r,\boldsymbol{V}) - \hat{m}_a(r,\boldsymbol{V}))\right] \\ &+ \mathbb{E}\left[\frac{\mathbb{I}(A=a)S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})\hat{g}_a(\boldsymbol{X})}(q_a(r,\boldsymbol{X}) - \hat{q}_a(r,\boldsymbol{X}))\right],\end{aligned}$$

where $\tilde{m}_a(r,\boldsymbol{V}) = \mathbb{E}[\hat{q}_a(r,\boldsymbol{X})]$. We can remove dependence on $\tilde{m}_a(r,\boldsymbol{V})$ by noting the second term can be rewritten as

$$\mathbb{E}\left[\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}(\tilde{m}_a(r,\boldsymbol{V}) - m_a(r,\boldsymbol{V}))\right] - \mathbb{E}\left[\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}(\hat{m}_a(r,\boldsymbol{V}) - m_a(r,\boldsymbol{V}))\right],$$

where we can then leverage the fact that

$$\mathbb{E}\left[\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}(\tilde{m}_a(r,\boldsymbol{V})-m_a(r,\boldsymbol{V}))\right]=\mathbb{E}\left[\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}(\hat{q}_a(r,\boldsymbol{X})-q_a(r,\boldsymbol{X}))\right].$$

Given this form for the second term, after re-arranging we can rewrite $\mathbb{E}(\chi_a(r,\hat{\eta})-\chi_a(r,\eta))$ as

$$\begin{aligned}
&\mathbb{E}(\chi_a(r,\hat{\eta})-\chi_a(r,\eta))\\
&=\mathbb{E}\left[\frac{(1-\hat{\kappa}(\boldsymbol{V}))(\hat{q}_a(r,\boldsymbol{X})-q_a(\boldsymbol{X}))}{\hat{\kappa}(\boldsymbol{V})}\left\{S-\frac{\mathbb{I}(A=a)S}{\hat{g}_a(\boldsymbol{X})}\right\}\right]\\
&+\mathbb{E}\left[\left\{(1-S)-\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}\right\}(\hat{m}_a(\boldsymbol{V})-m_a(\boldsymbol{V}))\right]\\
&=\mathrm{I}+\mathrm{II}.
\end{aligned}$$

Now, term I above can be bounded by noting

$$\begin{aligned}
\mathrm{I}&=\mathbb{E}\left[\frac{(1-\hat{\kappa}(\boldsymbol{V}))(\hat{q}_a(r,\boldsymbol{X})-q_a(r,\boldsymbol{X}))}{\hat{\kappa}(\boldsymbol{V})}\left\{S-\frac{\mathbb{I}(A=a)S}{\hat{g}_a(\boldsymbol{X})}\right\}\right]\\
&=\mathbb{E}\left[\frac{\mathbb{P}(S=1|\boldsymbol{X})(1-\hat{\kappa}(\boldsymbol{V}))(\hat{q}_a(r,\boldsymbol{X})-q_a(r,\boldsymbol{X}))}{\hat{\kappa}(\boldsymbol{V})\hat{g}_a(\boldsymbol{X})}\{\hat{g}_a(\boldsymbol{X})-g_a(\boldsymbol{X})\}\right]\\
&\leq\frac{1}{\varepsilon'}\mathbb{E}[(\hat{q}_a(r,\boldsymbol{X})-q_a(r,\boldsymbol{X}))\cdot(\hat{g}_a(\boldsymbol{X})-g_a(\boldsymbol{X}))]\\
&\leq\frac{1}{\varepsilon'}||\hat{q}_a(r,\boldsymbol{X})-q_a(r,\boldsymbol{X})||\cdot||\hat{g}_a(\boldsymbol{X})-g_a(\boldsymbol{X})||,
\end{aligned}$$

where $\varepsilon'>0$. Above, the third line holds by positivity conditions outlined at the beginning of the proof, while the fourth line holds by the Cauchy-Schwarz inequality.

Through similar logic, we can bound the second term by noting

$$\begin{aligned}
\mathrm{II}&=\mathbb{E}\left[\left\{(1-S)-\frac{S(1-\hat{\kappa}(\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}\right\}(\hat{m}_a(\boldsymbol{V})-m_a(\boldsymbol{V}))\right]\\
&=\mathbb{E}\left[\frac{(\hat{\kappa}(\boldsymbol{V})-\kappa(\boldsymbol{V}))(\hat{m}_a(r,\boldsymbol{V})-m_a(r,\boldsymbol{V}))}{\hat{\kappa}(\boldsymbol{V})}\right]\\
&\leq\frac{1}{\varepsilon'}||\hat{\kappa}-\kappa||\cdot||\hat{m}_a(r)-m_a(r)||
\end{aligned}$$

Notice I and II imply that

$$\begin{aligned}
\mathbb{E}(\chi_a(\hat{r}_{a,\alpha},\hat{\eta})-\chi_a(\hat{r}_{a,\alpha},\eta))&\leq\sup_r\frac{1}{\varepsilon'}||\hat{q}_a(r)-q_a(r)||\cdot||\hat{g}_a-g_a||+\sup_r||\hat{\kappa}-\kappa||\cdot||\hat{m}_a(r)-m_a(r)||\\
&=O_{\mathbb{P}}\left(\sup_r||\hat{q}_a(r)-q_a(r)||\cdot||\hat{g}_a-g_a||+\sup_r||\hat{\kappa}-\kappa||\cdot||\hat{m}_a(r)-m_a(r)||\right)
\end{aligned}$$

where we use the fact that by construction $||\hat{q}_a(\hat{r}_{a,\alpha})-q_a(\hat{r}_{a,\alpha})||\leq\sup_r||\hat{q}_a(r)-q_a(r)||$, analogously holding for $\hat{m}_a(r)$ Recalling $\mathbb{P}(Y(a)\in\hat{C}_a(\boldsymbol{V})|S=0)-1-\alpha=\mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha};\eta)]/\mathbb{P}(S=0)$ and the decomposition of $\mathbb{E}[\chi_a(\boldsymbol{O},\hat{r}_{a,\alpha}\ \eta)]$ yields the desired result.

Appendix B. Additional Experiments Details

B.1. Methods Implementation

We briefly provide additional information on the implementation of the naive DML estimator which ignores runtime confounding, and the weighted estimator explored throughout our numerical experiments. Details on specific training parameters are provided later in the section.

In implementing the naive DML estimator, we follow Algorithm 1, enforcing $\boldsymbol{X} = \boldsymbol{V}$. In the setting where one forces $\boldsymbol{X} = \boldsymbol{V}$ by ignoring runtime confounding, the EIC reduces to

$$\chi_a(r_{a,\alpha}, \boldsymbol{O}\,; \eta_a(r_{a,\alpha})) = (1-S)(q_a(r_{a,\alpha}, \boldsymbol{V}) - (1-\alpha)) + w_a(\boldsymbol{O})\{\mathbb{I}(R_a(Y, \boldsymbol{V}) \leq r_{a,\alpha}) - q_a(r_{a,\alpha}, \boldsymbol{V})\},$$

where $\tilde{w}_a(\boldsymbol{O}) = \frac{AS(1-\kappa(\boldsymbol{V}))}{\tilde{g}_a(\boldsymbol{V})\kappa(\boldsymbol{V})}$, where $\tilde{g}_a(\boldsymbol{V}) := \mathbb{P}(A = a|\boldsymbol{V}, S = 1)$. Intuitively, with this restriction we effectively have $m_a(r, \boldsymbol{V}) = q_a(r, \boldsymbol{V})$, canceling out middle term in the original EIC.

The weighted estimator is obtained through a split conformal prediction procedure which solves the estimating equation implied by Equation 4 on the calibration fold of source observations.

B.2. Numerical Experiments

In this section, we provide full details on the procedures used to generate data in our numerical and semi-synthetic experiments in 5.

B.2.1. DATA GENERATION

Our numerical experiments extend the setup considered in Coston et al. (2020). Letting $\boldsymbol{V} = (V_1, \ldots, V_{p_V})$ and $\boldsymbol{U} = (U_1, \ldots, U_{p_U})$

$$\begin{aligned}
V_k &\sim \mathcal{N}(0,1), \quad 1 \leq k \leq p_V \\
U_k &\sim \mathcal{N}(0,1), \quad 1 \leq k \leq p_U \\
Y(a) &= \mu(\boldsymbol{V}, \boldsymbol{U}) + \epsilon(\boldsymbol{V}, \boldsymbol{U}), \quad \mu(\boldsymbol{V}, \boldsymbol{U}) = \frac{k_V}{k_V + k_U}\left(\sum_{k=1}^{k_V} V_k + 2\sum_{k=1}^{k_U} U_i\right) \\
A &\sim \text{Bernoulli}(\pi(\boldsymbol{V}, \boldsymbol{U})), \qquad g(\boldsymbol{V}, \boldsymbol{U}) = \text{expit}\left(\frac{1}{\sqrt{k_V + k_U}}\left(\sum_{i=1}^{k_V} V_i - 2\sum_{i=1}^{k_U} U_i\right)\right) \\
S &\sim \text{Bernoulli}(\kappa(\boldsymbol{V})), \qquad \kappa(\boldsymbol{V}) = \text{expit}\left(b - \frac{1}{\sqrt{k_V}}\sum_{k=1}^{k_V} V_k\right)
\end{aligned}$$

where $\text{expit}(x) = \exp(x)/(1+\exp(x))$, $\epsilon(\boldsymbol{V}, \boldsymbol{U}) \sim N(0, \sqrt{|\mu(\boldsymbol{V}, \boldsymbol{U})|})$ $\boldsymbol{V} = (V_1, \ldots, V_{k_V})$, $\boldsymbol{U} = (U_1, \ldots, U_{k_U})$, $k_V \leq p_V$ and $k_U \leq p_U$. We choose b in $\kappa(\boldsymbol{V})$ to ensure $\mathbb{E}[\kappa(\boldsymbol{V})] = \mathbb{P}(S = 1) = 0.9$, achieving this numerically by simulating 1 million values of $\boldsymbol{V}$ outside of our main simulation.

Notably, source population membership is influenced by $\boldsymbol{V}$, generating covariate shift between the source and target populations. A and $Y(a)$ are both influenced by $\boldsymbol{V}$ and $\boldsymbol{U}$. To induce runtime confounding, we treat $\boldsymbol{U}$ as unobserved in the target population ($S = 0$).

We set $p_V = p_U = 15$, $k_V = 5$ and vary $k_U \in \{5, 10, 15\}$. This setup induces sparsity in the outcome, treatment and population models, while allowing us to investigate the impact of increasingly severe instances of runtime confounding.

We allow for covariate shift between $S = 0$ and $S = 1$ units by simulating S as a function of $\boldsymbol{V}$, extending the setup considered in Coston et al. (2020).

B.2.2. TRAINING DETAILS

Constructing prediction intervals among the three approaches considered requires estimation of g_a, κ, q_a and m_a. We additionally require estimation of μ_a and η_a when using absolute residual conformity scores, and estimation of $Q_{a,\alpha}$ when using quantile conformity scores.

We fit all of g_a, κ, q_a, m_a, with a stacked ensemble of random forests and Lasso models. We fit these ensemble learners with the `SuperLearner` package in R. Random forests are fit with the `ranger` package, using default hyperparameters specified by the `SL.ranger` SuperLearner library. Lasso models are fit with the `glmnet` package, similarly choosing default values specified by `SL.glmnet`. Both estimated treatment and source probability are trimmed to be within the interval $(0.025, 0.975)$ to avoid instability induced by large inverse propensity weights.

When using absolute residual conformity scores, we use the two-stage procedure proposed by Coston et al. (2020) and described in Section 3. In this setting, μ_a and η_a are fit with this same stacked ensemble with `SuperLearner`. When using methods which ignore runtime confounding, effectively $\mu_a = \eta_a$, meaning we only fit μ_a.

When using quantile conformity scores, we fit $Q_{a,\alpha}$ with weighted quantile forests, using the weights we propose in Proposition 1. We use weights of the form $\hat{w}_a(\boldsymbol{O})$, recalling $\hat{w}_a(\boldsymbol{O})$ is a function of $\hat{g}_a$ and $\hat{\kappa}$ fit according to the procedure above, and use the `ranger` package to implement the corresponding weighted quantile regression, specifying the same parameters as above. When using methods which ignore runtime confounding, we perform unweighted quantile regression.

B.3. Data Application

B.3.1. DATA GENERATION DETAILS

We emulate the data generating procedure employed by Lei and Candès (2021), additionally enforcing runtime confounding. We describe the procedure here, emphasizing that the data generation closely follows the procedure described in Lei and Candès (2021). Following Lei and Candès (2021), we split the ACIC data into two folds, Z_1 and Z_2, where $|Z_1| = 2079$ and $|Z_2| = 8312$.

To investigate varying degrees of runtime confounding, we consider three splits of $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{U})$:

1. **Severe**: $\boldsymbol{V} = (\texttt{X3}, \texttt{X4}, \texttt{X5}, \texttt{XC})$ and $\boldsymbol{U} = (\texttt{X1}, \texttt{X2}, \texttt{X5}, \texttt{C1}, \texttt{C3}, \texttt{S3})$
2. **Moderate**: $\boldsymbol{V} = (\texttt{X1}, \texttt{X2}, \texttt{X3}, \texttt{X4}, \texttt{X5}, \texttt{XC})$, $\boldsymbol{U} = (\texttt{C1}, \texttt{C2}, \texttt{C3}, \texttt{S3})$
3. **Mild**: $\boldsymbol{V} = (\texttt{X1}, \texttt{X2}, \texttt{X3}, \texttt{X4}, \texttt{X5}, \texttt{XC}, \texttt{S3}, \texttt{C1})$, $\boldsymbol{U} = (\texttt{C2}, \texttt{C3})$

On Z_1 and for each runtime confounding scenario we consider, we

- Fit $\hat{m}_0(\boldsymbol{X}) = \hat{\mathbb{E}}[Y(0)|\boldsymbol{X}]$ with the `randomForest` package.
- We fit $\hat{g}(\boldsymbol{X}) = \hat{\mathbb{P}}(A = 1|\boldsymbol{X})$ through the `randomForest` package, truncating $\hat{g}(\boldsymbol{X})$ to fall within 0.1 and 0.9
- We fit the 25% and 75% conditional quantile functions for $Y(0)$ and $Y(1)$ with the `grf` package, and let $\hat{r}_0(\boldsymbol{X})$ and $\hat{r}_1(\boldsymbol{X})$ denote the corresponding interquartile ranges

- We regress `Z` on $\boldsymbol{V}$ with the `randomForest` package, obtaining predictions $\hat{\kappa}(\boldsymbol{V})$. Treating the predicted probabilities as $\tilde{\kappa}(\boldsymbol{V})$, we produce $\kappa(\boldsymbol{V})$ by choosing b such that $\mathbb{E}[\text{expit}(b - \text{logit}(\tilde{\kappa}(\boldsymbol{V}))] = 0.9$.

We then let

$$Y_i(0) = \hat{m}_0(\boldsymbol{X}_i) + 0.5\hat{r}_0(\boldsymbol{X}_i)\varepsilon_{i0}, \quad Y_i(1) = \hat{m}_1(\boldsymbol{X}_i) + \tau(\boldsymbol{X}_i) + 0.5\hat{r}_1(\boldsymbol{X}_i)\varepsilon_{i1},$$

where ε_{ia} are iid $N(0,1)$ for $a = 0, 1$ and $\tau(\boldsymbol{X})$ is the CATE function defined in equation (1) of Carvalho et al. (2019). We then simulate data according to

$$\begin{aligned}
\boldsymbol{X} &\sim F_{Z_2} \\
A|\boldsymbol{X} &\sim \text{Bernoulli}(\hat{g}(\boldsymbol{X})) \\
Y_i(0) &= \hat{m}_0(\boldsymbol{X}_i) + 0.5\hat{r}_0(\boldsymbol{X}_i)\varepsilon_{i0} \\
Y_i(1) &= \hat{m}_1(\boldsymbol{X}_i) + \tau(\boldsymbol{X}_i) + 0.5\hat{r}_1(\boldsymbol{X}_i)\varepsilon_{i1}, \\
S &\sim \text{Bernoulli}(\hat{\kappa}(\boldsymbol{V})),
\end{aligned}$$

where F_{Z_2} is the empirical distribution of covariates in the held out split of data Z_2. Note that we enforce a runtime confounding scenario by simulating source population membership through S as a function of $\boldsymbol{V}$, extending the setup considered in Lei and Candès (2021).

In implementing our considered methods, nuisance functions, we train nuisance functions using the same learners considered for our numerical simulations.

Appendix C. Additional Numerical Experiment Results

C.1. Results Stratified by Treatment Level

Our main results pool coverage rates and average interval lengths for both $Y(1)$ and $Y(0)$ in the target population. Figure 4 reports coverage rates and interval lengths separately for $Y(1)$ and $Y(0)$. Qualitative results are similar. Average interval lengths are typically larger for $Y(1)$.

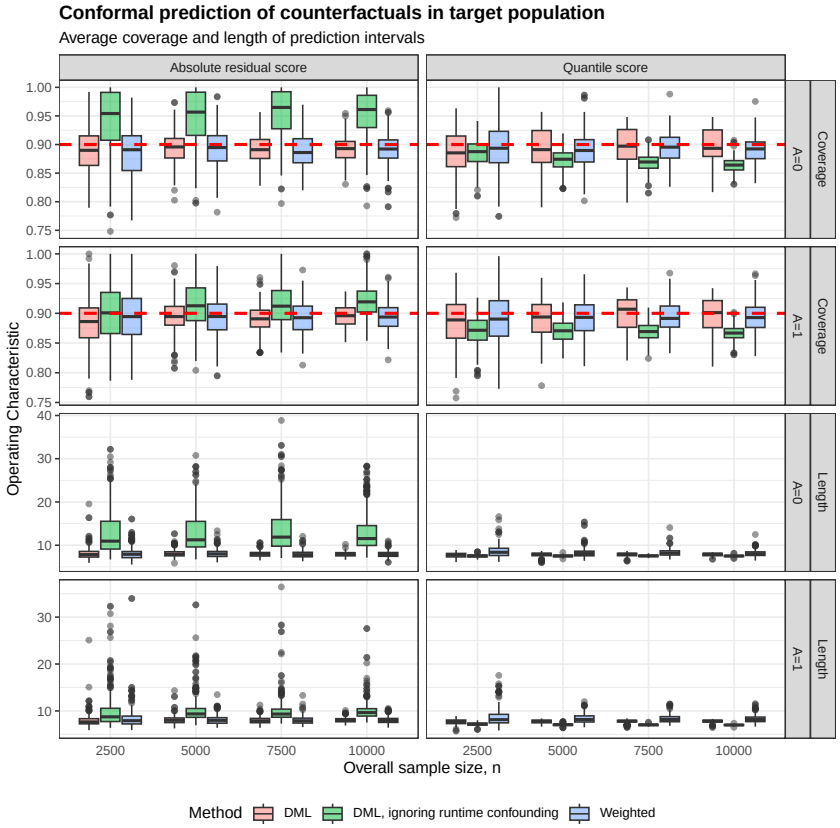


Figure 4: Performance of proposed methods stratified by counterfactual outcome.

C.2. Varying the Degree of Runtime Confounding

We report results stratified by treatment level a when varying the degree of true runtime confounders, controlled by k_V in Section B. Results remain qualitatively similar to our baseline scenario where $k_V = 10$.

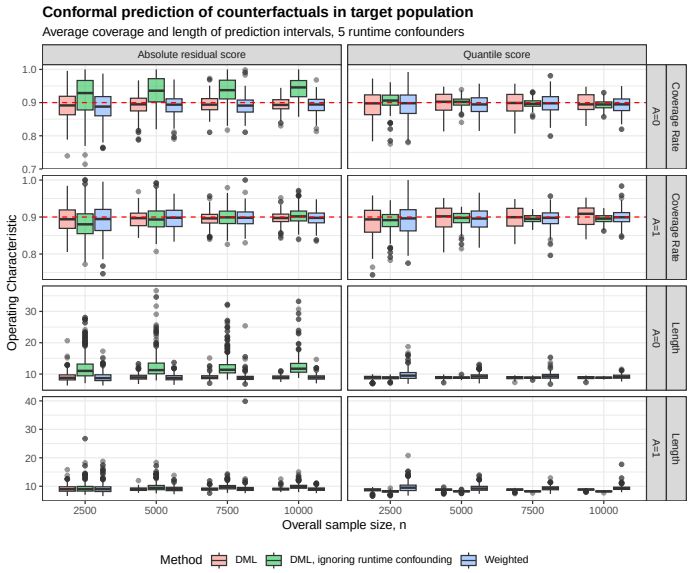


Figure 5: Performance of proposed methods, varying n and fixing the number of runtime confounders at 5.

C.3. Varying the Share of Target Population Data

Fixing $n = 5000$ and the number of runtime confounders at 10, we vary the share of source data $\mathbb{P}(S = 1)$, finding similar results across all shares considered.

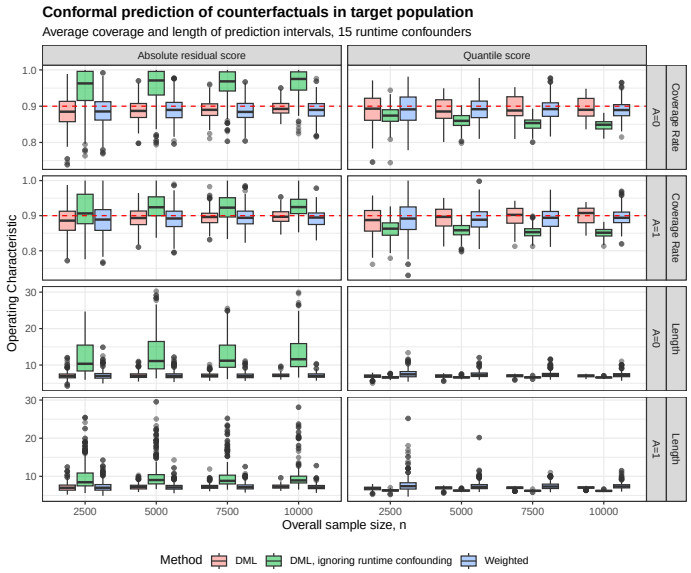


Figure 6: Performance of proposed methods, varying n and fixing the number of runtime confounders at 15.

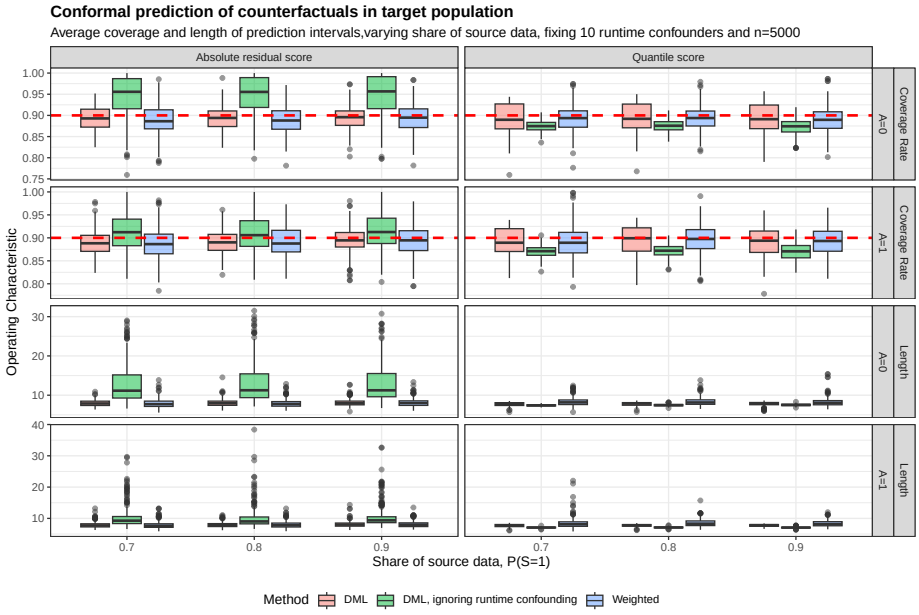


Figure 7: Performance of proposed methods when varying $\mathbb{P}(S = 1)$, fixing $n = 5000$ and the number of runtime confounders at 10.

Appendix D. Additional Semi-Synthetic Experiment Results

D.1. Baseline Results Stratified by Treatment Level

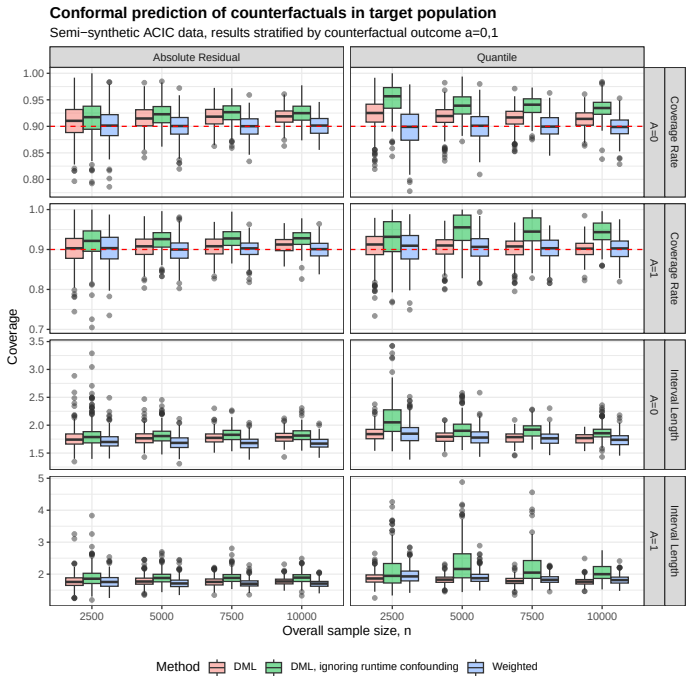


Figure 8: Performance of proposed methods on semi-synthetic ACIC data, varying n under the baseline moderate runtime confounding scenario.

D.2. Varying Degree of Runtime Confounding

In this section, we report the results obtained by repeating our semi-synthetic ACIC data exercise when the set of variables included in $\boldsymbol{V}$ varies as outlined in Appendix B.3. Intuitively, the fewer covariates available in $\boldsymbol{V}$, the greater the degree of runtime confounding. We see that the naive DML approach deteriorates in the severe runtime confounding scenario, often producing excessively wide intervals relative to the weighted and DML approaches.

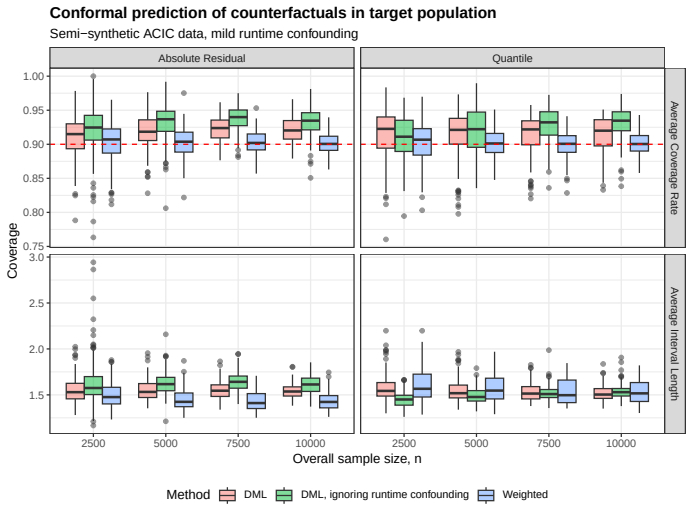


Figure 9: Performance of proposed methods on semi-synthetic ACIC data under the mild runtime confounding scenario.

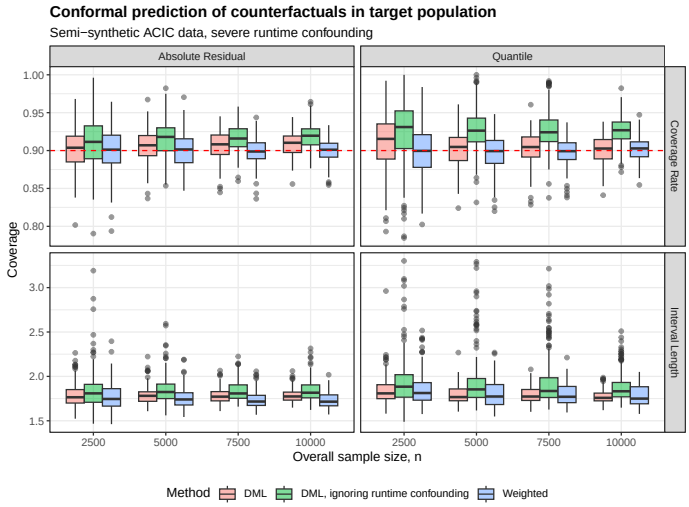


Figure 10: Performance of proposed methods on semi-synthetic ACIC data under the severe runtime confounding scenario.

Appendix E. Intervals for Individual Treatment Effects

While we fixate interest on the setting where A is a categorical random variable and interest lies in the construction of intervals for $Y(a)|S=0$ for generic a, a large set of work has covered the setting where $A \in \{0,1\}$ and interest lies in constructing intervals for target population individual treatment effects $Y(1)-Y(0)|S=0$.

Following Lei and Candès (2021), a straightforward approach to construct prediction intervals $\hat{C}_{\text{ITE}}(\boldsymbol{V})$ targeting the coverage result

$$\mathbb{P}(Y(1)-Y(0) \in \hat{C}_{\text{ITE}}(\boldsymbol{V})|S=0) = 1-\alpha$$

is to

1. Construct $1-\alpha/2$ level intervals $\hat{C}_1(\boldsymbol{V}) = (\hat{C}_1^{\text{L}}(\boldsymbol{V}), \hat{C}_1^{\text{U}}(\boldsymbol{V}))$ and $\hat{C}_0(\boldsymbol{V}) = (\hat{C}_0^{\text{L}}(\boldsymbol{V}), \hat{C}_0^{\text{U}}(\boldsymbol{V}))$ for $Y(1)$ and $Y(0)$, respectively, using Algorithm 1
2. Construct intervals of the form $\hat{C}_{\text{ITE}}(\boldsymbol{V}) = (\hat{C}_1^{\text{L}}(\boldsymbol{V}) - \hat{C}_0^{\text{U}}(\boldsymbol{V}), \hat{C}_1^{\text{U}}(\boldsymbol{V}) - \hat{C}_0^{\text{L}}(\boldsymbol{V}))$

Although easy to implement, the above approach will tend to produce excessively wide intervals. Alternatively, one can construct nested intervals as outlined in Lei and Candès (2021) and later extended to handle target-source covariate shift in a surrogate outcome setting by Gao et al. (2025). Although not the focus of this work, we briefly discuss the high-level procedure one can follow:

1. Within the source population, construct intervals $\hat{C}(\boldsymbol{X})$ which aim to satisfy

   $$\mathbb{P}(Y(1)-Y(0) \in \hat{C}(\boldsymbol{X})|S=1)$$

   To do this, suppose $C_a(\boldsymbol{X})$ satisfies $\mathbb{P}(Y(a) \in C_a(\boldsymbol{X})|S=1, A=1-a)$. Since A is observed for all units in the source population, one can construct ITE intervals in the source population of the form

   $$C(\boldsymbol{X}) = \begin{cases} Y - C_0(\boldsymbol{X}), & A \cdot S = 1, \\ C_1(\boldsymbol{X}) - Y, & (1-A) \cdot S = 1 \end{cases}$$

   The component intervals $C_a(\boldsymbol{X})$ can be constructed using the doubly-robust procedure proposed in Yang et al. (2024), where all of $\boldsymbol{X}$ can be used since $\boldsymbol{X}$ is available for all members of the source population.

2. Define a conformity score $R_C(C, \boldsymbol{V})$ with respect to the individual-level intervals $\hat{C}(\boldsymbol{X}_i)$ in the source population. Gao et al. (2025) provide recommendations for choices of scores, where here we restrict the scores to incorporate only $\boldsymbol{V}$ since $\boldsymbol{U}$ is unobserved in the target population

3. Target the $1-\gamma$ quantile of R_C in the target population, denoted r_γ which satisfies

   $$\mathbb{P}(R_C(C, \boldsymbol{V}) \leq r_\gamma | S=0) = 1-\gamma,$$

   noting under the earlier independence assumptions we will have r_γ additionally satisfies

   $$\mathbb{E}[\mathbb{P}(R_C(C, \boldsymbol{V}) \leq r_\gamma | S=1, \boldsymbol{V}) | S=0] = 1-\gamma.$$

Given the above identifying functional, one can construct doubly-robust estimators of r_γ using the approach outlined in Gao et al. (2025), forming intervals of the form $C_{\text{ITE}}(\boldsymbol{V}) = \{c : R_C(c, \boldsymbol{V}) \leq \hat{r}_\gamma\}$, who established the resulting intervals asymptotically satisfy

$$\mathbb{P}(Y(1) - Y(0) \in C_{\text{ITE}}(\boldsymbol{V}) | S = 0) \geq 1 - (\alpha + \gamma).$$

under standard regularity conditions. While the above procedure will yield intervals with the desired properties, we devote a formal implementation and study of the resulting intervals to future work.

Appendix F. Discussion of the Independence Assumption 4

As discussed in Section 2, it can be difficult to assess the plausibility of Assumption 4 in multi-source settings. In this Section, we briefly discuss recommendations for assessing the plausibility of this Assumption.

We begin by noting Assumption 4 is implied by the following two alternative Assumptions:

Assumption 5.a $Y(a) \perp\!\!\!\perp S|\boldsymbol{X}$

Assumption 5.b $\boldsymbol{U} \perp\!\!\!\perp S|\boldsymbol{V}$

To see this, assume for simplicity that the data are discrete and note that for any y, v, s

$$\begin{aligned}\mathbb{P}(Y(a)=y)|\mathbf{V}=v,S=s) &= \sum_u \mathbb{P}(Y(a)=y)\big|\mathbf{V}=v,\mathbf{U}=u,S=s)\mathbb{P}(\mathbf{U}=u|V=v,S=s) \\ &= \sum_u \mathbb{P}(Y(a)=y)\big|\mathbf{V}=v,\mathbf{U}=u)\mathbb{P}(\mathbf{U}=u|V=v),\end{aligned}$$

where the last display does not depend on s, implying $Y(a) \perp\!\!\!\perp S|\boldsymbol{V}$, which is exactly Assumption 4. Assumption 5.a can be viewed as a weaker version of Assumption 4 that conditions on the full set of covariate information, which in tandem with Assumption 3 implies that the set of covariates $\mathbf{X}$ that are sufficient to control for treatment-outcome confounding in the source population are additionally sufficient to render $Y(a)$ independent from S. Relatedly, Assumption 5.b implies there is no covariate shift in $\mathbf{U}$ across populations conditional on the always-observed $\mathbf{V}$, which may be plausible in settings where the source and target sites do not enroll.

While we believe it often easiest to assess the plausibility of Assumption 4 through the plausibility of both , since we rely on their implied condition $Y(a) \perp\!\!\!\perp S|\boldsymbol{V}$ for identification, we invoke this condition directly in the manuscript. In light of this alternative framing, we discuss examples in which we expect Assumption 4 to hold below, and provide example DAGs where Assumption 4 is violated in Figure 11.

Example Scenarios where Assumption 4 will be Plausible

To develop intuition for determining the plausibility of Assumption 4, consider a runtime confounding setting involving the treatment of acute ischemic stroke. Interest lies in forming counterfactual prediction intervals for the impact of different treatments A, (e.g. thrombectomy) on hospital length of stay among individuals receiving care from a target population hospital, using data from a separate hospital corresponding to the source population.

Suppose $\boldsymbol{V}$ collects baseline demographic characteristics and readily obtainable information including blood pressure, age, and NIH stroke scale. Further suppose $\boldsymbol{U}$ contains additional information which informs treatment decisions in the source population—such as cerebral blood flow—but is more resource-intensive to collect and in turn not readily available in the target population hospital. Assumption 5.a will be plausible if $\boldsymbol{V}$ and $\boldsymbol{U}$ explain away hospital-specific effects on length of stay, and Assumption 5.b will be plausible if the target and source population hospitals enroll

similar patient populations at baseline. Recall that these two Assumptions in turn imply the desired condition $Y(a) \perp\!\!\!\perp S|\boldsymbol{V}$.

Alternatively, Assumption 5.b may be less plausible if the target and source hospitals enroll patients with notably different baseline characteristics, and Assumption 5.a may be less plausible if features unmeasured in both sites but tied to hospital quality—such as staff size—meaningfully influence length of stay.

Figure 11: Two possible directed acyclic graphs consistent where Assumption 3 is satisfied but 4 is violated.