Conformal prediction for causal effects of continuous treatments

Maresa Schröder
LMU Munich
Munich Center for Machine Learning
maresa.schroeder@lmu.de

Dennis Frauen
LMU Munich
Munich Center for Machine Learning
frauen@lmu.de

Jonas Schweisthal
LMU Munich
Munich Center for Machine Learning
jonas.schweisthal@lmu.de

Konstantin Hess
LMU Munich
Munich Center for Machine Learning
k.hess@lmu.de

Valentyn Melnychuk
LMU Munich
Munich Center for Machine Learning
melnychuk@lmu.de

Stefan Feuerriegel
LMU Munich
Munich Center for Machine Learning
feuerriegel@lmu.de

Abstract

Uncertainty quantification of causal effects is crucial for safety-critical applications such as personalized medicine. A powerful approach for this is conformal prediction, which has several practical benefits due to model-agnostic finite-sample guarantees. Yet, existing methods for conformal prediction of causal effects are limited to binary/discrete treatments and make highly restrictive assumptions, such as known propensity scores. In this work, we provide a novel conformal prediction method for potential outcomes of continuous treatments. We account for the additional uncertainty introduced through propensity estimation so that our conformal prediction intervals are valid even if the propensity score is unknown. Our contributions are three-fold: (1) We derive finite-sample validity guarantees for prediction intervals of potential outcomes of continuous treatments. (2) We provide an algorithm for calculating the derived intervals. (3) We demonstrate the effectiveness of the conformal prediction intervals in experiments on synthetic and real-world datasets. To the best of our knowledge, we are the first to propose conformal prediction for continuous treatments when the propensity score is unknown and must be estimated from data.

1 Introduction

Machine learning (ML) for estimating causal quantities such as causal effects and the potential outcomes of treatments is nowadays widely used in real-world applications such as personalized medicine [12]. However, existing methods from causal ML typically focus on point estimates [e.g., 38, 40], which means that the uncertainty in the predictions is neglected and hinders the use of causal ML in safety-critical applications [12, 31]. As the following example shows, uncertainty quantification (UQ) of causal quantities is crucial for reliable decision-making.

Motivating example: Let us consider a doctor who seeks to determine the dosage of chemotherapy in cancer care. This requires estimating the tumor size in response to the dosage for a specific patient

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

profile. A point estimate will predict the *average* size of the tumor post-treatment, but it will neglect that chemotherapy is ineffective for some patients. In contrast, UQ will give a *range* of the tumor size that is to be expected post-treatment, so that doctors can assess the probability that the patients will actually benefit from treatment. This helps to understand the risk of a treatment being ineffective and can guide doctors to choose treatments that are effective *with large probability*.

A powerful method for UQ is ***conformal prediction*** (CP) [35, 39, 51]. CP provides model-agnostic and distribution-free, finite-sample validity guarantees for quantifying uncertainty. CP has been widely used for traditional, predictive ML [e.g., 4, 6, 18], where it has been shown to yield reliable prediction intervals in finite-sample settings (see Fig. 1). Recently, there have been works that adapt CP for estimating causal quantities (see Fig. 2 for an overview). Yet, existing methods for CP focus on binary or discrete treatments [e.g., 2, 27, 36], but *not* continuous treatments, which is our novelty.

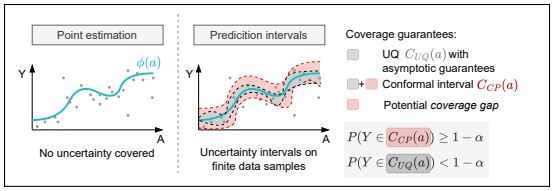

Figure 1: CP intervals on finite-sample data. UQ methods with asymptotic guarantees might suffer from under-coverage and are often *not* faithful. Thus, we aim at CP with finite-sample guarantees.

Adapting CP to causal quantities is *non-trivial* for two main reasons. **Challenge** (a): Intervening on the treatment induces a shift in the covariate distribution, specifically in the propensity score. As a result, the so-called *exchangeability assumption*, which is inherent to CP [51], is violated between the observational and interventional distribution[1], and because of this, standard CP intervals are not valid. Thus, we must later account for the distribution shift and derive *treatment-conditional guarantees*. **Challenge** (b): Assessing the aforementioned shift in the distribution requires information about the propensity score; yet, the propensity score is typically unknown. Hence, estimating the propensity score introduces *additional uncertainty*. However, incorporating the additional uncertainty in the overall CP intervals cannot be done in a simple plug-in manner, and it is highly non-trivial.

Unique to CP for effects of continuous treatments is a third challenge (c): data points with the same treatment value are rarely observed. Thus, we later employ smoothing to model the propensity shift.

In this paper, we develop a CP method for causal quantities, such as potential outcomes, of continuous treatments. Our method is designed to account for the additional uncertainty introduced during propensity estimation and is thus applicable to settings where the propensity score is unknown.

Our contributions:[2] (1) We propose a novel method for CP of causal quantities such as potential outcomes or treatment effects of continuous treatments. For this, we mathematically derive finite-sample prediction intervals for potential outcomes under known and unknown propensity functions. (2) We provide an algorithm for efficiently calculating the derived intervals. (3) We demonstrate the effectiveness of the derived CP intervals in experiments on multiple datasets.

2 Related Work

UQ for causal effects: Existing methods for UQ of causal quantities are often based on Bayesian methods [e.g., 1, 21, 22, 24]. However, Bayesian methods require the specification of a prior distribution based on domain knowledge and are thus neither robust to model misspecification nor generalizable to model-agnostic machine learning models. A common ad hoc method for computing uncertainty intervals is Monte Carlo (MC) dropout [14]. However, MC dropout yields approximations of the posterior distribution, which are *not* faithful [34].

Conformal prediction: CP [35, 39, 51] has recently received large attention for finite-sample UQ. For a prediction model ϕ trained on dataset $D_T = (X_i, Y_i)_{i=1,\dots,m}$ and a new test sample X_k, CP aims to construct a prediction interval $C(X_k)$ such that $P(Y_k \in C(X_k)) \geq 1-\alpha$ for some significance level α. We refer to [4] for an in-depth overview. Due to its strong finite-sample validity guarantees, CP is widely used for traditional, predictive ML with widespread applications such as in medical settings [57] or drug discovery [3, 10].

[1] By interventional we refer to the distribution after the intervention

[2] Code and data are available at our public GitHub repository: `https://github.com/m-schroder/ContinuousCausalCP`

Several extensions have been developed for CP. One literature stream focuses on CP with *marginal coverage* under distribution shifts between training and test data [e.g., 8, 11, 15, 16, 17, 18, 19, 36, 41, 48, 55]. Our setting later also involves a distribution shift due to the intervention on the treatment but differs from the latter in that the true distribution shift is unknown. Another literature stream constructs intervals *conditional* on the variables following the shifted distribution. Since, in general, exact conditional coverage has been proven impossible [35, 50], the works in this literature stream have two key limitations: (1) they only guarantee *approximate* conditional coverage [e.g., 5, 7, 35, 43]; or (2) they are restricted to specific data structures such as binary variables [e.g., 35, 50]. Because of that, none of the existing methods for marginal and conditional coverage can be applied to derive prediction intervals with finite-sample validity guarantees for causal quantities of continuous treatments.

Causal conformal prediction	Unknown propensity	Finite sample exact guarantees	Continuous treatment
e.g., Alaa et al. (2023), Wang et al. (2024)	✗	✓	✗
e.g., Jin et al. (2023), Jonkers et al. (2024)	✓	✗	✗
Lei and Candès (2021)	✓	✓	✗
Ours	✓	✓	✓

Figure 2: Key works on causal CP.

Conformal prediction for causal quantities: Only a few works focus on CP for causal quantities (see Fig. 2). Examples are methods aimed at off-policy learning [47, 58], conformal sensitivity analysis [56], or meta-learners for the conditional average treatment effect (CATE) [2, 27, 36, 52]. However, there are crucial differences to our setting: First, the existing works (a) assume that the propensity is *known* and thus achieve finite-sample coverage guarantees, or the existing works (b) focus on the easier task of giving *asymptotic* guarantees but then might suffer from under-coverage because of which the intervals are *not* faithful. Only Lei and Candès [36] provides finite-sample coverage guarantees under estimated propensity scores. However, all existing CP methods are designed for *binary* or *discrete* treatments. Applying such methods to discretized continuous treatments leads to ill-defined causal estimands. Therefore, none of the existing methods are applicable to our continuous treatment setting. We offer a detailed discussion in Supplement F.

Research gap: To the best of our knowledge, no work has provided prediction intervals with finite-sample validity guarantees for causal quantities of continuous treatments.

3 Problem formulation

Notation: We denote random variables by capital letters X with realizations x. Let P_X be the probability distribution over X. We omit the subscript whenever it is obvious from the context. For discrete X, we denote the probability mass function by $P(x) = P(X = x)$ and the conditional probability mass functions by $P(y \mid x) = P(Y = y \mid X = x)$ for a discrete random variable Y. For continuous X, $p(x)$ is the probability density function w.r.t. the Lebesgue measure.

Setting: Let the data $(X_i, A_i, Y_i)_{i=1,\dots,n}$ consisting of observed confounders $X \in \mathcal{X}$, a continuous treatment $A \in \mathcal{A}$, and an outcome $Y \in \mathcal{Y}$ be drawn *exchangeably* from the joint distribution P. Additionally, let a new sample of confounders X_{n+1} be drawn independently from the marginal distribution P_X. Throughout our work, we split the dataset into a proper training dataset $D_T = (X_i, A_i, Y_i)_{i=1,\dots,m}$, and a calibration dataset $D_C = (X_i, A_i, Y_i)_{i=m+1,\dots,n}$. Furthermore, let $\pi(a \mid x)$ define the generalized propensity score for treatment $A = a$ given $X = x$.

Throughout this work, we build upon the potential outcomes framework [44]. We denote the potential outcomes of a hard intervention a^* by $Y(a^*)$ and of a soft intervention $A^*(x) \sim \tilde{\pi}(a \mid x) = P_{A^* \mid X=x}$ by $Y(A^*(x))$.[3] We make three standard identifiability assumptions for causal effect estimation: positivity, consistency, and unconfoundedness [e.g., 2, 27]. Finally, we consider an arbitrary machine learning model ϕ to predict the potential outcomes. Hence, we define the outcome prediction function as $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, $\phi(X, A) \mapsto Y$. We assume the dose-response curve to be sufficiently Hölder-smooth. This is common in settings with continuous treatments [e.g., 40, 45].

[3] Interventions are characterized by two classes: hard (structural) and soft (parametric) interventions. Hard interventions directly affect the treatment by setting it to a specific value and removing the edge in the graph (as in the do-operator). Soft interventions do not change the structure of the graph but affect the conditional distribution of the treatment given the confounders. All interventions affect the propensity score, but the mathematical consequences are different. For soft interventions, the new (interventional) propensity is a function of the original propensity. For hard interventions, the new propensity is given by the Dirac-delta function.

Our objective: In this work, we aim to derive *conformal prediction intervals* $C(X_{n+1}, \lozenge)$ for the prediction of a potential outcome $Y_{n+1}(\lozenge)$ of a new data point under either hard, $\lozenge = a^*$, or soft intervention, $\lozenge = A^*(X_{n+1}) \sim \tilde{\pi}(a \mid X_{n+1})$. The derived intervals are called *valid* for any new exchangeable sample X_{n+1} with non-exchangeable intervention $\lozenge$, i.e., for $\lozenge \in \{a^*, A^*(X_{n+1})\}$ and significance level $\alpha \in (0, 1)$

$$P(Y_{n+1}(\lozenge) \in C(X_{n+1}, \lozenge)) \geq 1 - \alpha. \tag{1}$$

Of note, our CP method can be used with an arbitrary ML model ϕ to predict the potential outcomes.

In CP, the interval C is constructed based on so-called *non-conformity scores* [51], which capture the performance of the prediction model ϕ. For example, a common choice for the non-conformity score is the residual of the fitted model $s(X, A, Y) = |Y - \phi(X, A)|$, which we will use throughout our work. For ease of notation, we define $S_i := s(X_i, A_i, Y_i)$.

Why is CP for causal quantities non-trivial? There are two main reasons. First, coverage guarantees of CP intervals essentially rely on the exchangeability of the non-conformity scores. However, intervening on treatment A shifts the propensity function and, therefore, induces a shift in the covariates (X, A) ($\rightarrow$ Challenge (a)). Formally, we have a *propensity shift* in which the intervention $\lozenge$ shifts the propensity function $\pi(a \mid x)$ to either a Dirac-delta distribution of the hard intervention, $\delta_{a^*}(a)$, or to the distribution of the soft intervention, $\tilde{\pi}(a \mid x)$, without affecting the outcome function $\phi(x, a)$. As a result, the test data sample under $\lozenge$ does *not follow the same distribution* as the train and calibration data, i.e., the exchangeability assumption is violated.

Second, the propensity score π is commonly *unknown* in observational data and, therefore, must be estimated, which introduces additional uncertainty that one must account for when constructing CP intervals ($\rightarrow$ Challenge (b)). Crucially, existing coverage guarantees [e.g., 50, 48] do *not* hold in our setting. Instead, we must derive new intervals with valid *coverage under propensity shift*.

In the following, we address the above propensity shift by performing a calibration conditional on the shift induced by the intervention, which allows us then to yield valid prediction intervals with significance level $(1 - \alpha)$ for potential outcomes of a specific hard or soft intervention. We emphasize that the extension to intervals for causal effects is straightforward, in that one combines the intervals for each potential outcome under a certain treatment and without treatment, so that eventually arrives at CP intervals for the individual treatment effect (ITE). Details are in Supplement A.

4 CP intervals for potential outcomes of continuous treatments

Recall that intervening on test data breaks the necessary exchangeability assumption, i.e., the guaranteed coverage of at least $(1 - \alpha)$. Therefore, we now construct CP intervals where we account for a (potentially unknown) propensity shift in the test data induced by the intervention.

Scenarios: In our derivation, we distinguish two different scenarios(see Fig. 3):

① **Known propensity score** (see Section 4.1): If the propensity score in the observational data is known, it means that the treatment policy is known. Then, we aim to update the policy by increasing/decreasing the treatment by a value Δ_A, i.e., $A^*(X) = A + \Delta_A$.
Example: A doctor prescribes a medication to a new patient. Instead of prescribing the same dosage as he would have prescribed to a similar patient in the past, the doctor is interested in the potential health outcome when increasing (or decreasing) the original dosage by an amount Δ_A.

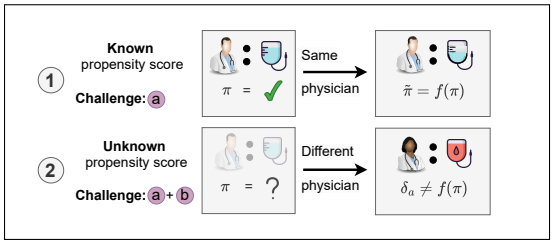

Figure 3: Use cases of the two scenarios: ① The new assignment is a function of the original policy (i.e., soft intervention). ② The policy in the dataset is unknown. The new assignment cannot be expressed as a function of the original policy (i.e., hard intervention).

② **Unknown propensity score** (see Section 4.2): In observational data, the propensity score is unknown. Therefore, we usually assess the effect of hard interventions, i.e., a^*. Here, we face

additional uncertainty from propensity score estimation ($\rightarrow$ Challenge (b)).[4] *Example:* In our running example, a patient sees a doctor who has never prescribed the respective medication and thus will base the decision on observational data (electronic health records), which was collected under a different, unknown treatment policy (e.g., from another physician). Thus, the prescribed intervention (dosage) cannot be expressed in terms of the policy in the observational data.

In our derivations, we make use of the following two mathematical tools. First, we define the *propensity shift*. Formally, it is the shift between the observational and interventional distributions, P and $\tilde{P}$, in terms of the tilting of the propensity function by a non-negative function f. Hence, we have

$$\tilde{\pi}(a \mid x) = \frac{f(a,x)}{\mathbb{E}_P[f(A,X)]}\pi(a \mid x). \tag{2}$$

for some f with $\mathbb{E}_P[f(X,A)] > 0$ and $a \in \mathcal{A}, x \in \mathcal{X}$.

Second, our CP method will build upon ideas from so-called split conformal prediction [39, 51], yet with crucial differences. In our methods, the calibration step differs from the standard procedure in that we *conditionally calibrate* the non-conformity scores depending on the tilting function f to achieve marginal coverage for the interventional – and thus shifted – data.

High-level outline: Our derivation in Sections 4.1 and 4.2 proceed as follows. Following [18, 42], we reformulate split conformal prediction as an augmented quantile regression. Let S_i represent the non-conformity score of the sample (X_i, A_i, Y_i) for $i = m+1, \ldots, n$ of the calibration dataset and $S_{n+1} = S$ an imputed value for the unknown score of the new sample. We define

$$\hat{\theta}_S := \underset{\theta \in \mathbb{R}}{\arg\min} \frac{1}{n-m} \left(\sum_{i=m+1}^{n} l_\alpha(\theta, S_i) + l_\alpha(\theta, S) \right), \tag{3}$$

where

$$l_\alpha(\theta, S) := (\alpha - \mathbf{1}_{[\theta - S < 0]})(\theta - S) = \begin{cases} (1-\alpha)(S-\theta), & \text{if } S \geq \theta, \\ \alpha(\theta - S), & \text{if } S < \theta. \end{cases} \tag{4}$$

Of note, $\hat{\theta}_S$ is an estimator of the $(1-\alpha)$-quantile of the non-conformity scores [32, 46]. Using $\hat{\theta}_S$, we construct the CP interval with the desired coverage $(1-\alpha)$. However, the interval is only valid for exchangeable data. Quantile regression might yield non-unique solutions that can depend on the indices of the scores [18], so we later restrict the analysis to solvers invariant to the data ordering.[5]

4.1 Scenario 1: Known propensity score

We first consider scenario (1) with known propensity scores. Here, existing CP intervals are not directly applicable due to the shift from old to new propensity ($\rightarrow$ Challenge (a)). For our derivation, we need the following lemma building upon and generalizing the intuition presented above.

Lemma 4.1 ([18])**.** *Let $\mathcal{F}$ define a finite-dimensional function class that includes the function f characterizing the shift in the (potentially unknown) propensity function π (see Eq. 2). Define the distribution-shift-calibrated $(1-\alpha)$-quantile of the non-conformity scores as*

$$\hat{g}_S(X_{n+1}) := \underset{g \in \mathcal{F}}{\arg\min} \frac{1}{n-m} \left(\sum_{i=m+1}^{n} l_\alpha(g(X_i), S_i) + l_\alpha(g(X_{n+1}), S) \right) \tag{5}$$

for an imputed guess S of the $(n+1)$-th non-conformity score S_{n+1}. The prediction interval

$$C(X_{n+1}) := \{y \mid S_{n+1}(y) \leq \hat{g}_{S_{n+1}(y)}(X_{n+1})\} \tag{6}$$

for the true S_{n+1} given a realization of $Y_{n+1} = y$ satisfies the desired coverage guarantee under all distribution shifts $f \in \mathcal{F}$, i.e.,

$$P_f(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha. \tag{7}$$

[4] Throughout our main paper, we focus on the setting of hard interventions. In some cases, it might also be of interest to perform soft interventions on the estimated propensity score. We provide derivations for this setting in Supplement A.

[5] We note that commonly used solvers, such as interior point solvers, are invariant to the data ordering.

Building upon Lemma 4.1, we derive our first main result in Theorem 4.2. We define the finite-dimensional function class of interest as $\mathcal{F} := \{\theta \frac{\pi(a+\Delta_A|x)}{\pi(a|x)} \mid \theta \in \mathbb{R}^+\}$. It is easy to verify that all $f \in \mathcal{F}$ represent the desired propensity shift to $\tilde{\pi}(a \mid x) = \pi(a + \Delta_A \mid x)$ as defined in Eq. 2.

However, note that the optimization problem in Eq. 5 requires knowledge about the *true* or *optimal* imputed scores S_{n+1}. Directly solving the problem in this form would require running it for all possible imputed values $S \in \mathbb{R}$, i.e., an infinite amount of times. As a remedy, we exploit the *strong duality property* and present our results in terms of a dual problem formulation.

Theorem 4.2 (Conformal prediction intervals for known baseline policy)**.** *Consider a new datapoint with $X_{n+1} = x_{n+1}$, $A_{n+1} = a_{n+1}$, and $A^*(X_{n+1}) = a^* = a_{n+1} + \Delta_A$. Let $\eta^S = \{\eta^S_{m+1}, \dots, \eta^S_{n+1}\} \in \mathbb{R}^{n-m}$ be the optimal solution to*

$$\begin{aligned} &\max_{\eta_i, i=m+1,\dots,n+1} \min_{\theta>0} \sum_{i=m+1}^{n} \eta_i \left(S_i - \theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)} \right) + \eta_{n+1} \left(S - \theta \frac{\pi(a^* \mid x_{n+1})}{\pi(a_{n+1} \mid x_{n+1})} \right) \\ &\textit{s.t.} \qquad -\alpha \leq \eta_i \leq 1 - \alpha, \quad \forall i = m+1, \dots, n+1, \end{aligned} \tag{8}$$

for an imputed unknown $S_{n+1} = S$. Furthermore, let S^* be defined as the maximum S s.t. $\eta^S_{n+1} < 1 - \alpha$. Then, the prediction interval

$$C(x_{n+1}, a^*) := \{y \mid S_{n+1}(y) \leq S^*\} \tag{9}$$

satisfies the desired coverage guarantee

$$P(Y(A^*(X_{n+1})) \in C(X_{n+1}, A^*(X_{n+1}))) \geq 1 - \alpha. \tag{10}$$

Proof. We provide a full proof in Supplement D.2. Here, we briefly outline the underlying idea of the proof. First, we show that the function class $\mathcal{F}$ indeed satisfies Eq. equation 2 for the intervention $A^*(X) = A + \Delta_A$, and we then rewrite Eq. equation 5 as a convex optimization problem. Next, we exploit the strong duality property. We optimize over the corresponding dual problem to receive a dual prediction set with equal coverage probability. Finally, we derive S^* from the dual prediction set to construct C_{n+1} and prove the overall coverage guarantee. □

4.2 Scenario 2: Unknown treatment policy

If the underlying treatment policy is unknown, the only possible intervention is a hard intervention a^*. As described above, measuring the induced propensity shift is non-trivial due to two reasons: (i) The propensity model needs to be estimated, which introduces additional uncertainty affecting the validity of the intervals ($\rightarrow$ Challenge (b)). (ii) The density function corresponding to a hard intervention is given by the Dirac delta function

$$\delta_{a^*}(a) := \begin{cases} 0, & \text{for} \quad a \neq a^*, \\ \infty, & \text{for} \quad a = a^*, \end{cases} \tag{11}$$

which hinders a direct adaptation of Theorem 4.2 due to the inherent discontinuity of the improper function. Hence, we make the following assumption on the propensity estimator.

Assumption 1. *The estimation error of the propensity function $\hat{\pi}(a \mid x)$ is bounded in the sense that, for all $i = 1, \dots, n+1$, there exists $M > 0$ such that*

$$c_{a_i} := \frac{\hat{\pi}(a_i \mid x_i)}{\pi(a_i \mid x_i)} \in \left[\tfrac{1}{M}, M\right]. \tag{12}$$

Under Assumption 1, the distribution shift induced by the intervention is then defined as

$$\tilde{\pi}(a \mid x) = \frac{\delta_{a^*}(a)}{\hat{\pi}(a \mid x)} \frac{\hat{\pi}(a \mid x)}{\pi(a \mid x)} \pi(a \mid x) = c_a \frac{\delta_{a^*}(a)}{\hat{\pi}(a \mid x)} \pi(a \mid x) = \frac{f(a, x)}{\mathbb{E}_P[f(A, X)]} \pi(a \mid x), \tag{13}$$

for a suitable function f. We further formulate $\delta_{a^*}(a)$ in terms of a Gaussian function as

$$\delta_{a^*}(a) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a - a^*)^2}{2\sigma^2} \right). \tag{14}$$

This motivates the following lemma. Therein, we specify the class $\mathcal{F}$ of tilting functions f that represent the distribution shift induced by the hard intervention a^*.

Lemma 4.3. *For $\sigma > 0$, we define*

$$f(a, x) := \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp(-\frac{(a-a^*)^2}{2\sigma^2})}{\hat{\pi}(a \mid x)} \tag{15}$$

with $\mathbb{E}_P[f(A, X)] = 1$. Furthermore, we define the finite-dimensional function class $\mathcal{F}$

$$\mathcal{F} := \left\{ \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp(-\frac{(a-a^*)^2}{2\sigma^2})}{\hat{\pi}(a \mid x)} \;\middle|\; 0 < \sigma, \frac{1}{M} \leq c_a \leq M \right\}. \tag{16}$$

Then, $f(a, x) \in \mathcal{F}$ for all $c_a \in [\frac{1}{M}, M]$ and $\sigma \to 0$. As a result, the distribution shift

$$\tilde{\pi}(a \mid x) = \lim_{\sigma \to 0} \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp(-\frac{(a-a^*)^2}{\sigma^2})}{\hat{\pi}(a \mid x)} \pi(a \mid x) \tag{17}$$

can be represented in terms of Eq. equation 2 through functions $f \in \mathcal{F}$.

Proof. See Supplement D.1. □

Following the motivation in scenario ①, we thus aim to estimate the $(1-\alpha)$-quantile of the non-conformity scores under the distribution shift in Lemma 4.1. We can reformulate this problem as

$$\begin{aligned} \min_{\sigma>0, \frac{1}{M} \leq c_a \leq M} \quad & \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i \\ \text{s.t.} \quad & S_i - \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a_i-a^*)^2}{2\sigma^2}\right)}{\hat{\pi}(a_i \mid x_i)} - u_i + v_i = 0, \quad \forall i = m+1 \ldots, n+1, \\ & u_i, v_i \geq 0, \quad \forall i = m+1 \ldots, n+1 \end{aligned} \tag{18}$$

for the imputed score $S_{n+1} = S$. As the score is unknown, computing the CP interval would require solving equation 18 for all $S \in \mathbb{R}$, yet which is computationally infeasible. Before, we exploited properties of the dual optimization problem and the Lagrange multipliers of the convex problem in Theorem 4.2 to efficiently compute the CP intervals. However, the present non-convex problem does not automatically allow for the same simplifications. Instead, we now present a remedy for efficient computation of the CP intervals in the following lemma. We prove Lemma 4.4 in Supplement D.

Lemma 4.4. *The problem equation 18 is Type-I invex and satisfies the linear independence constraint qualification (LICQ).*

Lemma 4.4 allows us to derive properties of the present non-convex optimization problem in terms of the Karush-Kuhn-Tucker (KKT) conditions. For this, we note that the fulfillment of the LICQ serves as a sufficient regularity condition for the KKT to hold at any (local) optimum of equation 18. Combined with the Type-I invexity of the objective function and the constraints, the KKT conditions are not only necessary but also sufficient for a global optimum. As a result, we can employ the KKT conditions at the optimal values[6] σ^* and c_a^* to derive coverage guarantees of our CP interval in a similar fashion as in Theorem 4.2. We thus arrive at the following theorem to provide CP intervals for the scenario with unknown propensity scores.

Theorem 4.5 (Conformal prediction intervals for unknown propensity scores)**.** *Let $u^S = \{u_{m+1}^S, \ldots, u_{n+1}^S\}, v^S = \{v_{m+1}^S, \ldots, v_{n+1}^S\} \in \mathbb{R}^{n-m}$, $\sigma^S, c_a^S \in \mathbb{R}$ be the optimal solution to*

$$\begin{aligned} \min_{\sigma>0, \frac{1}{M} \leq c_a \leq M} \quad & \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i \\ \textit{s.t.} \quad & S_i - \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a_i-a^*)^2}{2\sigma^2}\right)}{\hat{\pi}(a_i \mid x_i)} - u_i + v_i = 0, \quad \forall i = m+1 \ldots, n+1 \\ & u_i, v_i \geq 0, \quad \forall i = m+1 \ldots, n+1 \end{aligned} \tag{19}$$

[6]We provide an interpretation of the optimal values in Supplement G. Therein, we further discuss the implications of the proposed kernel smoothing of $\delta_{a^*}(a)$.

for an imputed unknown $S_{n+1} = S$. Let S^* be defined as the maximum S s.t. $v_{n+1}^S > 0$. Then,

$$C(X_{n+1}, a^*) := \{y \mid S_{n+1}(y) \leq S^*\} \tag{20}$$

satisfies the desired coverage guarantee

$$P(Y(a^*) \in C(X_{n+1}, a^*)) \geq 1 - \alpha. \tag{21}$$

Proof. See Supplement D.3. □

In certain applications, it might be beneficial to fix σ to a small value σ_0 to approximate $\delta_{a^*}(a)$ though a soft intervention and only construct the CP interval through optimizing over c_a. We present an alternative theorem for this case in Supplement A.

We now use Thm. 4.5 to present an algorithm for computing CP intervals of potential outcomes from continuous treatment variables under unknown propensities in Alg. 1. We present a similar algorithm for scenario ① with known propensities and discuss the computational complexity in Supplement B.

5 Experiments

Baselines: As we have discussed above, there are *no* baselines that directly compute prediction intervals with finite-sample guarantees for potential outcomes of continuous treatments. Therefore, we compare our method against MC dropout [14] and deep ensemble methods [33]. Yet, we again emphasize that MC dropout is an ad hoc method with poor approximations of the posterior, which is known to give unfaithful intervals [34]. Furthermore, we report the empirical coverage achieved by intervals from a Gaussian process regression (GP). By doing so, we consider a method that assesses the underlying aleatoric uncertainty. Additionally, we compare our method against the naive vanilla CP and the method by [36] for binary treatments in Supplements C and G.

Implementation: All methods are implemented with ϕ as a multi-layer perceptron (MLP) and an MC dropout regularization of rate 0.1. Crucially, we use the *identical* MLP for both our CP method and MC dropout. Hence, all performance gains must be attributed to the coverage guarantees of our conformal method. In the MC dropout baseline, the uncertainty intervals are computed via Monte Carlo sampling. In scenario ②, we perform conditional density estimation by conditional normalizing flows [49]. Implementation and training details are in Supplement G.

Performance metrics: We evaluate the methods in terms of whether the prediction intervals are *faithful* [e.g., 21]. That is, we compute whether the *empirical coverage* of the prediction intervals surpasses the threshold of $1 - \alpha$ for different significance levels $\alpha \in \{0.05, 0.1, 0.2\}$. Additionally, we report the width of the resulting intervals in Supplement H.

5.1 Datasets

Synthetic datasets: We follow common practice and evaluate our methods using synthetic datasets [e.g., 2, 25]. Due to the fundamental problem of causal inference, counterfactual outcomes are never observable in real-world datasets. Synthetic datasets enable us to access counterfactual outcomes and thus to benchmark methods in terms of whether the computed intervals are faithful. Additionally, we perform experiments on the semi-synthic TCGA dataset in Supplement C. We hereby show the applicability of our method to high-dimensional real-world data in a controlled environment.

Medical dataset: We demonstrate the applicability of our CP method to medical datasets on the MIMIC-III dataset [26]. MIMIC-III contains de-identified health records from patients admitted to critical care units at large hospitals. Our goal is to predict patient outcomes in terms of blood pressure when treated with a different duration of mechanical ventilation. We use 8 confounders from medical practice (e.g., respiratory rate, hematocrit). Overall, we consider 14,719 patients, split into train (60%), validation (10%), calibration (20%), and test (10%) sets. Details are in Supplement G.

5.2 Results for synthetic datasets

We consider two synthetic datasets with different propensity scores and outcome functions. Dataset 1 uses a step-wise propensity function and a concave outcome function. Dataset 2 is more complex and uses a Gaussian propensity function and oscillating outcome functions. Both datasets contain a single discrete confounder, a continuous treatment, and a continuous outcome.

By choosing low-dimensional datasets, we later render it possible to plot the treatment–response curves so that one can inspect the prediction intervals visually. (We later also show that our method scales to high-dimensional settings as part of the real-world dataset.) Details about the data generation are in Supplement G.

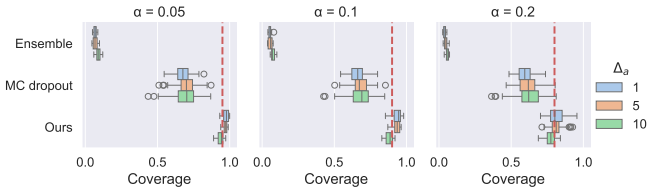

Figure 4: Comparison of *faithfulness* on dataset 1 across 50 runs. Larger values are better. For each α, the plots show how often the empirical intervals contain the true outcome. Intervals should ideally yield a coverage of $1 - \alpha$ (red line).

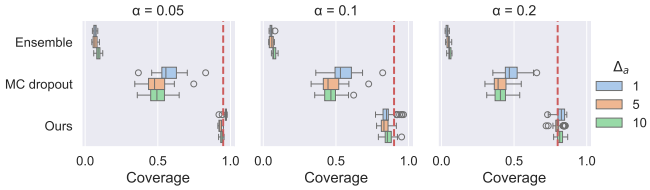

Figure 5: Comparison of *faithfulness* on dataset 2 across 50 runs. Larger values are better.

We evaluate the faithfulness of our CP intervals. On each dataset, we analyze the performance of the intervals in the presence of various soft interventions $\Delta_a \in \{1, 5, 10\}$ and hard interventions $a^* \in \{7x, 10x\}$ for each $X = x$. We average the empirical coverage across 50 runs with random seeds. The results are in Fig. 4 and Fig. 5. Additionally, we report the empirical coverage of the GP.

We make the following observations. First, our CP intervals comply with the targeted significance level α and are therefore faithful. Second, both MC dropout and the deep ensemble method have a considerably lower coverage, implying that the intervals are *not* faithful. This is in line with the literature, where MC dropout is found to produce poor approximations of the posterior [34]. In particular, the ensemble method is highly unfaithful. Thus, we will not consider this baseline in all the following experiments. Third, our method has only a small variability in terms of empirical coverage, whereas the empirical coverage of MC dropout varies greatly. This corroborates the robustness of our method. Fourth, the results are consistent for both datasets. Fifth, the GP is only able to capture the true potential outcome in the prediction intervals for small distribution shifts ($\Delta = 1$) on dataset 2. However, the empirical coverage is extremely low: $\alpha = 0.05$: 0.125; $\alpha = 0.1$: 0.125; $\alpha = 0.2$: 0.0833. This aligns with our expectations, as the aleatoric uncertainty in our experiments is low. Therefore, the GP intervals are small (average width of 0.1293) and barely valid after the intervention. In sum, this demonstrates our CP method's effectiveness.

Data	Intervention	Coverage $\alpha = 0.05$	Coverage $\alpha = 0.1$	Coverage $\alpha = 0.2$
1	$a^* = 7x$	**1.00** / 0.19	**0.90** / 0.13	**0.83** / 0.11
	$a^* = 10x$	**1.00** / 0.28	**0.91** / 0.23	**0.88** / 0.11
2	$a^* = 7x$	**1.00** / 0.02	**0.94** / 0.02	**0.85** / 0.02
	$a^* = 10x$	**1.00** / 0.08	**0.84** / 0.07	**0.83** / 0.07

Table 1: Coverage of the intervals from our CP method / MC dropout for various hard interventions a^* and significance levels α. Intervals with coverage $\geq 1 - \alpha$ are considered faithful.

Table 1 presents the empirical coverage of the intervals from our CP method vs. MC dropout across different α and hard interventions a^*. We observe that our CP intervals are effective and achieve the intended coverage. In contrast, MC dropout does not provide faithful intervals. Our findings are again in line with the literature, where MC dropout is found to produce poor

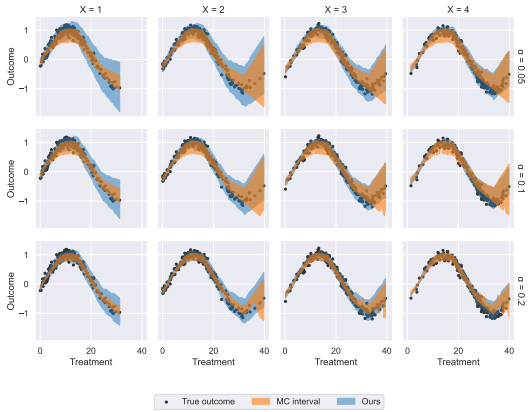

Figure 6: CP intervals for multiple significance levels α for Dataset 1 with intervention $\Delta = 5$.

approximations of the posterior and thus might provide poor coverage [34]. We present further results in Supplement H.

Insights: We plot the intervals across different levels α and covariates X (Fig. 6), allowing us to inspect the intervals visually. We observe that the intervals behave as expected: they become sharper with increasing significance level α. We further see that our CP intervals are slightly wider (see details in Supplement H), yet this is intended as it ensures that the intervals are faithful. Our CP intervals (blue) generally include the true outcome. In contrast, the intervals from MC dropout (orange) often do *not* include the true outcome (e.g., see the bottom row Fig. 6) and are thus *not* faithful.

5.3 Results for the MIMIC dataset

Recall that numerically evaluating causal inference methods on real-world data is not possible due to the fundamental problem of causal inference. Therefore, we provide insights on the MIMIC dataset in Fig. 7. We compare the CP intervals of two male and two female patients of differing ages when treated with increasing duration of mechanical ventilation. Our intervals show higher uncertainty in treatment regions rarely included in the training data (high medium to high treatments). The intervals given by MC-dropout are narrower, which suggests lower coverage, confirming the effectiveness of our proposed method. This finding aligns with our observation from the synthetic datasets.

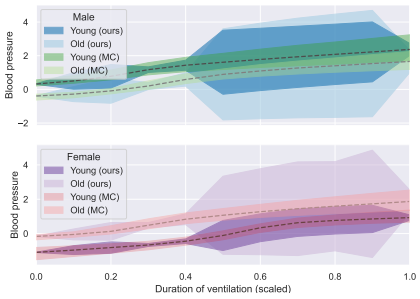

Figure 7: CP intervals for potential outcomes of increasing duration of mechanical ventilation for 4 exemplary patients.

6 Discussion

Limitations: As with any other method, our UQ method has limitations that offer opportunities for future research. Our method relies on the quality of the propensity estimator. Although we incorporate estimation errors when constructing intervals, poorly estimated propensities could potentially lead to wide prediction intervals. We acknowledge that our intervals are conservative for point interventions and for segments of the output space with limited calibration data, implying that a representative calibration dataset is essential for the performance of our method. As for all CP methods, the use of sample splitting may reduce data efficiency. Furthermore, we note that the optimization procedure can be computationally expensive for large CATE vectors.

Broader impact: Our method makes a significant step toward UQ for potential outcomes and, thus, toward *reliable* decision-making. We provided strong theoretical and empirical evidence that our prediction intervals are valid. To this end, our method fills an important demand for using causal ML in medical practice and other safety-critical applications with limited data.

Considerations for practical application: First, our method is designed for single continuous treatments with a univariate outcome, which is common in dosing (e.g., determining the dosage of chemotherapy/insulin/hypertension drugs). In *randomized controlled trials* (when the propensity score is known), we can directly apply Theorem 4.2 on top of the trained outcome prediction model ϕ to construct our CP intervals at target coverage level α. In observational studies, when the propensity score is unknown, we require the practitioner to have sufficient prior knowledge to set a bound M on the propensity estimation error. The practitioner should choose this bound with care and rather in a conservative way to prevent undercoverage. Then, the CP intervals can be derived based on Theorem 4.5. We emphasize again that our intervals do not suffer from undercoverage due to limited data, as CP guarantees are valid for *any sample size* and our method is kept fully *model agnostic*, enabling the practitioner to make reliable judgements based on limited data and an ML model of choice.

Conclusion: We presented a novel conformal prediction method for potential outcomes of continuous treatments with finite-sample guarantees. Our method extends naturally to treatment effects. A key strength of our method is that the intervals are valid under distribution shifts introduced by the treatment assignment, even if the propensity score is unknown and has to be estimated.

Acknowledgements

This paper is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

[1] A. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[2] A. Alaa, Z. Ahmad, and M. van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[3] J. Alvarsson, S. Arvidsson McShane, U. Norinder, and O. Spjuth. Predicting with confidence: Using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1): 42–49, 2021.

[4] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[6] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 2023.

[7] T. T. Cai, M. Low, and Z. Ma. Adaptive confidence bands for nonparametric regression functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 109:1054–1070, 2014.

[8] M. Cauchois, S. Gupta, A. Ali, and J. C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 119(548):3033–3044, 2024.

[9] Z. Chen, R. Guo, J.-F. Ton, and Y. Liu. Conformal counterfactual inference under hidden confounding. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[10] M. Eklund, U. Norinder, S. Boyer, and L. Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):117–132, 2015.

[11] C. Fannjiang, S. Bates, A. N. Angelopoulos, J. Listgarten, and M. I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences of the United States of America*, 119(43):e2204569119, 2022.

[12] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, and M. van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

[13] D. Frauen, F. Imrie, A. Curth, V. Melnychuk, S. Feuerriegel, and M. van der Schaar. A neural framework for generalized causal sensitivity analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

[14] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

[15] A. Gendler, T.-W. Weng, L. Daniel, and Y. Romano. Adversally robust conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

[16] S. Ghosh, Y. Shi, T. Belkhouja, Y. Yan, J. Doppa, and B. Jones. Probabilistically robust conformal prediction. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

[17] I. Gibbs and E. J. Candès. Adaptive conformal inference under distribution shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[18] I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025.

[19] L. Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

[20] M. A. Hanson and Mond. Necessary and sufficient conditions in constrained optimization. *Mathematical Programming*, 37(1):51–58, 1987.

[21] K. Hess, V. Melnychuk, D. Frauen, and S. Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *International Conference on Learning Representations (ICLR)*, 2024.

[22] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[23] G. W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[24] A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

[26] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[27] J. Jonkers, J. Verhaeghe, G. van Wallendael, L. Duchateau, and S. van Hoecke. Conformal Monte Carlo meta-learners for predictive inference of individual treatment effects. *arXiv preprint*, arXiv:2402.04906, 2024.

[28] N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[29] E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

[30] D. Kivaranovic, R. Ristl, M. Posch, and H. Leeb. Conformal prediction intervals for the individual treatment effect. *arXiv preprint*, arXiv:2006.01474, 2020.

[31] T. Kneib, A. Silbersdorff, and B. Säfken. Rage against the mean: A review of distributional regression approaches. *Econometrics and Statistics*, 26(C):99–123, 2023.

[32] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[33] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[34] L. Le Folgoc, V. Baltatzis, S. Desai, A. Devaraj, S. Ellis, O. E. M. Manzanera, A. Nair, H. Qiu, J. Schnabel, and B. Glocker. Is MC dropout Bayesian? *arXiv preprint*, arXiv:2110.04286, 2021.

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

[36] L. Lei and E. 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.

[37] L. Nagalapatti, A. Iyer, Abir, and S. Sarawagi. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. In *Conference on Artificial Intelligence (AAAI)*, 2024.

[38] L. Nie, M. Ye, Q. Liu, and D. Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations (ICLR)*, 2021.

[39] H. Papadopoulos. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, 2002.

[40] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Conference on Artificial Intelligence (AAAI)*, 2020.

[41] A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[42] Y. Romano, E. Patterson, and E. J. Candès. Conformalized quantile regression. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[43] Y. Romano, M. Sesia, and E. J. Candès. Classification with valid and adaptive coverage. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[44] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[45] J. Schweisthal, D. Frauen, V. Melnychuk, and S. Feuerriegel. Reliable off-policy learning for dosage combinations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[46] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

[47] M. F. Taufiq, J.-F. Ton, R. Cornish, Y. W. Teh, and A. Doucet. Conformal off-policy prediction in contextual bandits. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[48] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[49] B. L. Trippe and R. E. Turner. Conditional density estimation with bayesian normalising flows. *arXiv preprint*, arXiv:1802.04908, 2018.

[50] V. Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning (ACML)*, 2012.

[51] V. Vovk, A. Gammerman, and G. Shafer. *On-line compression modeling I: conformal prediction. In: Algorithmic Learning in a Random World*. Springer, Boston, 2005.

[52] B. Wang, F. Li, and M. Yu. Conformal causal inference for cluster randomized trials: model-robust inference without asymptotic approximations. *arXiv preprint*, arXiv:2401.01977, 2024.

[53] S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann. MIMIC-Extract. In *Conference on Health, Inference, and Learning (CHIL)*, 2020.

[54] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[55] Y. Yang, A. K. Kuchibhotla, and E. T. 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.

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

[57] X. Zhan, Z. Wang, M. Yang, Z. Luo, Y. Wang, and G. Li. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement*, 158: 107588, 2020.

[58] Y. Zhang, C. Shi, and S. Luo. Conformal off-policy prediction. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are later stated as theorems and proofed in the Appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in the end of the main paper as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All results are proven in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discusses the implementation details in ´the Appendix. Furthermore, the paper includes a link to the code for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides a link to an anonymized GitHub repository containing all code necessary for reproducing the results in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: A short section on the experimental setup is provided in the main paper. More details can be found in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the mean and standard deviation for the empirical coverage.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational complexity and compute resources are stated in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the broader societal impact of the contribution in the end of the main paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not pose the risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All original owners of code and (real-world) datasets are stated in the paper as well as the GitHub repository.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The usage of LLMs is not a standard component of the core methods in this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional theoretical results

A.1 Calculating prediction intervals for further causal quantities and differences

We presented a method for calculating conformal prediction intervals for potential outcomes of continuous treatments. In the following, we show how the intervals can be combined to yield valid prediction intervals for further causal quantities, such as the individual treatment effect (ITE) γ_i of treatment a:

$$\gamma_i(a) := Y_i(a) - Y_i(0). \tag{22}$$

Here, we consider the setting in which the non-conformity score is chosen to be the absolute residual.

Lemma A.1. *Let S_a^* and S_0^* denote the optimal imputed non-conformity scores S_{n+1} for treatment a and no treatment at significance level $1 - \frac{\alpha}{2}$ for $\alpha \in (0, 1)$, respectively. Furthermore, let*

$$C^+ := \phi(x_i, a) + S_a^* - \phi(x_i, 0) + S_0^*, \tag{23}$$

$$C^- := \phi(x_i, a) - S_a^* - \phi(x_i, 0) - S_0^*. \tag{24}$$

Then the interval $C_\gamma(X_i, a) := [C^-, C^+]$ contains the ITE γ_i with probability $1 - \alpha$.

Proof. Let $\varepsilon_i(a)$ be the estimation error of the potential outcome, i.e.

$$\varepsilon_i(a) := Y_i(a) - \phi(x_i, a). \tag{25}$$

We can rewrite the coverage guarantee of the original conformal prediction intervals for the potential outcome $Y(a)$ as

$$P(Y_i(a) \in C(X_i, a)) = P(|\varepsilon_i(a)| \leq S_a^*) \geq 1 - \frac{\alpha}{2}. \tag{26}$$

Now observe that

$$P(\gamma_i(a) \in C_\gamma(x_i, a)) = P((Y_i(a) - Y_i(0)) \in C_\gamma(x_i, a)) \tag{27}$$

$$= P((Y_i(a) \geq C^- + Y_i(0)) \wedge (Y_i(a) \leq C^+ + Y_i(0))) \tag{28}$$

$$= P((\varepsilon_i(a) \geq \varepsilon_i(0) - (S_a^* + S_0^*)) \wedge (\varepsilon_i(a) \leq \varepsilon_i(0) + (S_a^* + S_0^*))) \tag{29}$$

$$= P(|\varepsilon_i(a) - \varepsilon_i(0)| \leq S_a^* + S_0^*) \tag{30}$$

$$\geq P(|\varepsilon_i(a)| + |\varepsilon_i(0)| \leq S_a^* + S_0^*). \tag{31}$$

Thus, it follows directly that

$$P(\gamma_i(a) \in C_\gamma(X_i, a)) \geq 1 - \alpha. \tag{32}$$

□

A.2 Alternative scenario 2: Fixing an approximation of $\delta_{a^*}(a)$

In Section 4.2, we formulated the unknown propensity shift in terms of

$$\delta_{a^*}(a) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a - a^*)^2}{\sigma^2}\right) \tag{33}$$

and minimized over σ and c_a in Theorem 4.5 to construct the CP intervals. However, in certain applications, it might be beneficial to control the spread of the approximation of $\delta_{a^*}(a)$ through fixing σ to a small value σ_0 and performing the soft intervention $\tilde{\pi}(a \mid x) = \frac{c_a}{\sqrt{2\pi}\sigma_0} \frac{\exp\left(-\frac{(a-a^*)^2}{\sigma_0^2}\right)}{\hat{\pi}(a|x)}$. In this case, the resulting optimization problem is a convex problem similar to Theorem 4.2. We present the alternative optimization problem below.

Theorem A.2 (Alternative for Theorem 4.5: Conformal prediction intervals for unknown propensity scores)**.** *Let a new datapoint be given with $X_{n+1} = x_{n+1}$ and $A_{n+1} = a_{n+1}$. Let $\eta^S = \{\eta_1^S, \ldots, \eta_{n+1}^S\} \in \mathbb{R}^{n+1}$ be the optimal solution to*

$$\max_{\substack{\eta_i, \\ i=1,\ldots,n+1}} \min_{\frac{1}{M} \leq c_a \leq M} \sum_{i=1}^{n} \eta_i \left(S_i - \frac{c_a}{\sqrt{2\pi}\sigma_0} \frac{\exp\left(-\frac{(a_i - a^*)^2}{\sigma_0^2}\right)}{\hat{\pi}(a_i \mid x_i)}\right) + \eta_{n+1}\left(S - \frac{c_a}{\sqrt{2\pi}\sigma_0} \frac{1}{\hat{\pi}(a_i \mid x_{n+1})}\right)$$

$$\textit{s.t.} \qquad -\alpha \leq \eta_i \leq 1 - \alpha, \quad \forall i = 1, \ldots, n+1, \tag{34}$$

for an imputed unknown $S_{n+1} = S$. Furthermore, let S^* be defined as the maximum S s.t. $\eta^S_{n+1} < 1 - \alpha$. Then, the prediction interval

$$C(x_{n+1}, a^*) := \{y \mid S_{n+1}(y) \leq S^*\} \tag{35}$$

satisfies the desired coverage guarantee

$$P(Y(a^*) \in C(X_{n+1}, a^*) \geq 1 - \alpha, \tag{36}$$

where with a slight abuse of notation $Y(a^*)$ denotes the potential outcome under the soft intervention $\tilde{\pi}$ above.

Proof. The statement follows from Theorem 4.5. □

A.3 Soft-interventions on estimated propensities

In the main paper, we presented algorithms for constructing prediction intervals for soft interventions if the propensity function is known and hard interventions if it is unknown. These are arguably the most common scenarios in practice. However, in some cases, one might also be interested in the effect of soft interventions on estimated propensity scores [e.g., **?**]. Therefore, we present an alternative theorem for calculating conformal prediction intervals under soft interventions with estimated propensity scores below.

Theorem A.3 (Conformal prediction intervals for soft interventions with unknown propensity scores)**.** *Let a new datapoint be given with $X_{n+1} = x_{n+1}$ and $A_{n+1} = a_{n+1}$. Furthermore, let $\hat{\pi}$ denote the estimated propensity score with estimation error bounded by $[\frac{1}{M}, M]$, $M > 0$. The soft intervention is represented by the shift given through $\Delta \in \mathbb{R}$. Let $\eta^S = \{\eta^S_1, \ldots, \eta^S_{n+1}\} \in \mathbb{R}^{n+1}$ be the optimal solution to*

$$\begin{aligned} &\max_{\substack{\eta_i, \\ i=1,\ldots,n+1}} \min_{\frac{1}{M} \leq c_a \leq M} \sum_{i=1}^{n} \eta_i \left(S_i - \frac{c_a \hat{\pi}(a_i + \Delta \mid x_i)}{\hat{\pi}(a_i \mid x_i)} \right) + \eta_{n+1} \left(S - \frac{c_a \hat{\pi}(a_{n+1} + \Delta \mid x_{n+1})}{\hat{\pi}(a_{n+1} \mid x_{n+1})} \right) \\ &\textit{s.t.} \qquad -\alpha \leq \eta_i \leq 1 - \alpha, \quad \forall i = 1, \ldots, n+1, \end{aligned} \tag{37}$$

for an imputed unknown $S_{n+1} = S$. Furthermore, let S^* be defined as the maximum S s.t. $\eta^S_{n+1} < 1 - \alpha$. Then, the prediction interval

$$C(x_{n+1}, a^*) := \{y \mid S_{n+1}(y) \leq S^*\} \tag{38}$$

satisfies the desired coverage guarantee

$$P(Y(a^*) \in C(X_{n+1}, a^*) \geq 1 - \alpha, \tag{39}$$

where with a slight abuse of notation $Y(a^*)$ denotes the potential outcome under the soft intervention represented by Δ.

Proof. The statement follows from Theorem 4.2 and Theorem 4.5. □

B Algorithm

We now use Thm. 4.5 to present an algorithm for computing CP intervals of potential outcomes from continuous treatment variables under unknown propensities in Alg. 1. We present a similar algorithm for scenario ① with known propensities and discuss the computational complexity in below.

We make the following comments: In our algorithm, an optimization solver is used to calculate v_{n+1} according to Theorem 4.5. The specific choice of the solver is left to the user. In our experiments in Section 5, we perform the optimization via interior point methods. Further, the overall goal of our algorithm is to find the optimal imputed non-conformity score S^* such that the coverage guarantees hold. It can be implemented through suitable iterative search algorithms.

Algorithm 1: Algorithm for computing CP intervals of potential outcomes of continuous interventions under unknown propensities.

Input: Calibration data $(X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}$, new sample X_{n+1} and intervention a^*, significance level α, prediction model ϕ, propensity estimator $\hat{\pi}$, assumed error bound M, error tolerance ε, optimization solver

Output: CP interval C_{n+1} for a new test sample

1 $S_{\text{up}} \leftarrow \max\{\max_{i=m+1,\dots,n} S_i, 1\}$; $S_{\text{low}} \leftarrow \min\{\min_{i=m+1,\dots,n} S_i, -1\}$;

/* Calculate v_{n+1}^{up}, v_{n+1}^{low} */

2 $v_{n+1}^{\text{up}} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, M, S_{\text{up}})$;

3 $v_{n+1}^{\text{low}} \leftarrow \text{solver}((X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, M, S_{\text{low}})$;

/* Iterative search for S^* */

4 **while** $v_{n+1}^{up} > 0$ **do**

5 $\quad S_{\text{up}} \leftarrow 2S_{up}$;

6 $\quad v_{n+1}^{\text{up}} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, M, S_{\text{up}})$;

7 **end**

8 **while** $v_{n+1}^{\text{low}} = 0$ **do**

9 $\quad S_{\text{low}} \leftarrow 0.5S_{\text{low}}$;

10 $\quad v_{n+1}^{\text{low}} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, M, S_{\text{low}})$;

11 **end**

12 $S^* \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

13 **while** $S_{\text{up}} - S_{\text{low}} > \varepsilon$ **do**

14 $\quad v_{n+1}^{S^*} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, M, S^*)$;

15 $\quad$ **if** $v_{n+1}^{S^*} > 0$ **then**

16 $\quad\quad S_{\text{low}} \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

17 $\quad$ **end**

18 $\quad$ **else**

19 $\quad\quad S_v \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

20 $\quad$ **end**

21 $\quad S^* \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

22 **end**

/* Compute $C(X_{n+1}, a^*)$ */

23 **return** $C(X_{n+1}, a^*) = \{y \mid S_{n+1}(y) \leq S^*\}$

Algorithm explanation: Theorem 4.5 requires knowledge of the optimal imputed non-conformity score S^*. Since it is unknown beforehand, we implement Algorithm 1 as a binary search algorithm. We find suitable upper and lower bounds for S^* (until line 8). To validate that the bounds are indeed smaller/larger than the optimal non-conformity score, we observe the corresponding dual value ν_{n+1} (Theorem 4.5). After finding valid bounds, we start searching for S^* via standard binary search, continuously increasing the lower and decreasing the upper bound. When the difference between the bounds is less than a specified error tolerance ε, we stop and take S^* to be the mean of the interval. In every iteration (when calling 'solver'), we make use of the optimization in Theorem 4.5 and check whether the dual value fulfills the requirement in the Theorem.

Below, we state a second algorithm that is applicable if the propensity score is known. In this case, a convex optimization solver can be used.

Algorithm 2: Algorithm for computing CP intervals of potential outcomes of continuous interventions under known propensities.

Input: Calibration data $(X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}$, new sample X_{n+1} and soft intervention $A^*(X_{n+1})$, significance level α, prediction model ϕ, error tolerance ε, optimization solver

Output: CP interval C_{n+1} for a new test sample

$S_{\text{up}} \leftarrow \max\{\max_{i=m+1,\dots,n} S_i, 1\}$; $S_{\text{low}} \leftarrow \min\{\min_{i=m+1,\dots,n} S_i, -1\}$;

/* Calculate η^{up}_{n+1}, η^{low}_{n+1}, where η is the optimal solution to Eq. (8). */

$\eta^{\text{up}}_{n+1} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, A^*(X_{n+1}), \alpha, S_{\text{up}})$;

$\eta^{\text{low}}_{n+1} \leftarrow \text{solver}((X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, A^*(X_{n+1}), \alpha, S_{\text{low}})$;

/* Iterative search for S^* */

while $\eta^{up}_{n+1} < 1-\alpha$ **do**

  $S_{\text{up}} \leftarrow 2S_{up}$;

  $\eta^{\text{up}}_{n+1} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, A^*(X_{n+1}), \alpha, S_{\text{up}})$;

end

while $v^{\text{low}}_{n+1} >= 1-\alpha$ **do**

  $S_{\text{low}} \leftarrow 0.5S_{\text{low}}$;

  $\eta^{\text{low}}_{n+1} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, A^*(X_{n+1}), \alpha, S_{\text{low}})$;

end

$S^* \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

while $S_{\text{up}} - S_{\text{low}} > \varepsilon$ **do**

  $\eta^{S^*}_{n+1} \leftarrow \text{solver}(\phi, \hat{\pi}, (X_i, A_i, Y_i)_{i\in\{m+1,\dots,n\}}, X_{n+1}, a^*, \alpha, S^*)$;

  if $\eta^{S^*}_{n+1} < 1-\alpha$ **then**

    $S_{\text{low}} \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

  end

  else

    $S_{up} \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

  end

  $S^* \leftarrow \frac{S_{\text{up}}+S_{\text{low}}}{2}$;

end

/* Compute $C(X_{n+1}, A^*(X_{n+1}))$ */

return $C(X_{n+1}, A^*(X_{n+1})) = \{y \mid S_{n+1}(y) \leq S^*\}$

Computational complexity: The complexity of running our algorithms depends heavily on the employed optimization solver with complexity $\sigma_s(n_c)$ (e.g., polynomial complexity for suitable[7] convex solvers) and the size of the calibration dataset n_c. This might become costly for large-scale calibration datasets in practice. The outer algorithm has a time complexity of at most $O(\log(\frac{S_{up}-S_{low}}{\varepsilon}) + 1)$. Overall, our algorithm has a fixed complexity of $O(\log(\frac{S_{up}-S_{low}}{\varepsilon})\sigma_s(n_c))$. The complexity of deriving intervals through MC-dropout or ensemble methods depends, however, on the number of MC samples or models, respectively. The latter thus scales with the precision of the intervals, which might be difficult to control. Furthermore, we emphasize that MC intervals are generally *not* faithful and therefore *not* directly comparable. While the theoretical runtime of CP may exhibit more complex scaling behavior (e.g., cubic or non-linear), our empirical results demonstrate that CP scales well in practice. In our work, we use optimization as a tool to provide conformal prediction intervals. Future research should focus on developing more efficient optimization algorithms for this task.

Note on the stability of the optimization algorithm: In Scenario 1, the only source of potential instability can be a very low propensity score in low-overlap regions of the covariate space. This, however, is only a problem if $\pi(a|x) << \pi(a+\Delta|x)$. We can thus consider this unlikely in practice. For example, consider a patient who would be treated with a dosage of 10mg of some medication, which is prescribed in the range from 0 to 50mg. A practitioner is likely to be interested in the effect of an increase of the dosage to 15mg (i.e., $\Delta = 5$) in contrast to an increase to 50mg. Furthermore, it is reasonable to assume the propensity function to be locally smooth. Therefore, the instability of $\pi(a|x) << \pi(a+\Delta|x)$ is unlikely to occur.

[7] A suitable solver refers to a solver designed to handle the constrained convex problem in our theorem

In Scenario 2, we could additionally face an underflow of $\exp(-\frac{(a_i-a)^2}{2\sigma^2})$ or a blow-up of the prefactor $(\sigma\hat{\pi}(x_i))^{-1}$. As a remedy, we can reparameterize the problem to work in the log domain and only exponentiate after subtracting a stable (maximum) constant across all datapoints. This additionally prevents the Jacobian/Hessian of the constraints from becoming nearly singular or wildly varying, causing Newton-type steps to blow up or stall.

C Semi-synthetic experiments

To underline the effectiveness of our method, we perform additional experiments on the semi-synthetic TCGA dataset. The Cancer Genome Atlas (TCGA) dataset [54] consists of a comprehensive and diverse collection of gene expression data. The data was collected from patients with different cancer types. In our experiment, we consider the gene expression measurements of the 4,000 genes with the highest variability, which we employ as our features X. The study cohort consisted of a total of 9659 patients. We model a continuous treatment based on the sum of the 10 covariates with the highest variance and assign a treatment effect that is constant in the sum of the covariates. Specifically, we model the treatment to follow a normal distribution centered at 100*sum of the 10 covariate values, and the outcome to follow a normal distribution centered at the sum of the treatment and the covariate sum times 100.

As in the main paper, we construct CP intervals for different interventions and confidence levels α. We state the empirical coverage of our method in Table 2 as well as the coverage of the intervals returned by the ensemble method and MC dropout below. The prediction performance of the trained model on the hold-out test dataset is reported. We find that our method is highly effective.

Table 2: Coverage of the intervals from our CP method as well as the ensemble method and MC dropout on the TCGA dataset. We report the mean followed by the standard deviation in apprentices.

		Confidence level		
Intervention	Method	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
$\Delta = 0.5$	Ensemble	0.0640 (0.0445)	0.0600 (0.0379)	0.0480 (0.0412)
	MC	0.8280 (0.0795)	0.8160 (0.0783)	0.8040 (0.0741)
	Ours	0.9680 (0.0324)	0.8920 (0.0391)	0.8040 (0.0741)
$\Delta = 1.0$	Ensemble	0.0880 (0.0483)	0.0680 (0.0371)	0.0520 (0.0348)
	MC	0.8560 (0.0612)	0.8520 (0.0614)	0.8400 (0.0657)
	Ours	0.9733 (0.0377)	0.9500 (0.0500)	0.7667 (0.0618)
$\Delta = 1.5$	Ensemble	0.0720 (0.0411)	0.0680 (0.0412)	0.0640 (0.0389)
	MC	0.7880 (0.0815)	0.8040 (0.0925)	0.8280 (0.0786)
	Ours	0.9400 (0.0438)	0.8920 (0.0699)	0.8200 (0.0619)

We observe that our method consistently achieves the desired coverage. To evaluate the usefulness of our intervals, we also report the interval width in Table 3 below. The range of the outcomes was 2.0.

Table 3: Width of the intervals from our CP method on the TCGA dataset. We report the mean followed by the standard deviation in apprentices.

	Confidence level		
Intervention	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
$\Delta = 0.5$	0.1003 (0.0331)	0.0843 (0.0252)	0.0683 (0.0169)
$\Delta = 1.0$	0.1017 (0.0420)	0.0877 (0.0349)	0.0642 (0.0200)
$\Delta = 1.5$	0.0930 (0.0272)	0.0822 (0.0260)	0.0674 (0.0180)

Comparison to the CP methods for binary treatments [36]:

To investigate whether simple weighted CP methods for causal tasks on binary treatments sufficiently address the non-exchangeability due to the distribution shift induced by the intervention, we compare our method with the method by Lei and Candès [36]. Of note, both methods fulfill the coverage guarantees. However, when comparing the interval width, we see that our method is superior: Our intervals have an average width of 0.6255 (sd = 0.1714). In contrast, the intervals on the binarized treatment obtained through the method by Lei and Candès [36] have an average width of 3.2876 (sd = 0.5587). Hence, the intervals by our method are by far more informative.

D Proofs

D.1 Proofs of the supporting lemmas

In the following, we prove Lemma 4.3 and Lemma 4.4 from our main paper.

Proof of Lemma 4.3 Recall the definition of the hard intervention

$$\tilde{\pi}(a \mid x) = \delta_{a^*}(a) = \frac{\delta_{a^*}(a)}{\hat{\pi}(a \mid x)} \frac{\hat{\pi}(a \mid x)}{\pi(a \mid x)} \pi(a \mid x), \tag{40}$$

where

$$\delta_{a^*}(a) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a - a^*)^2}{\sigma^2} \right). \tag{41}$$

Under Assumption 1, we have

$$\frac{\hat{\pi}(a \mid x)}{\pi(a \mid x)} =: c_a \in [\frac{1}{M}, M] \tag{42}$$

for some $M > 0$ and all a, x. Then

$$\lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a-a^*)^2}{\sigma^2}\right)}{\hat{\pi}(a \mid x)} \frac{1}{M} \leq \pi(a^*|x) \leq \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a-a^*)^2}{\sigma^2}\right)}{\hat{\pi}(a \mid x)} M. \tag{43}$$

Therefore, the distribution shift induced by the hard intervention can be represented as

$$f(a, x) = \lim_{\sigma \to 0} \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a-a^*)^2}{\sigma^2}\right)}{\hat{\pi}(a \mid x)} \in \mathcal{F} := \left\{ \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a-a^*)^2}{\sigma^2}\right)}{\hat{\pi}(a \mid x)} \;\middle|\; 0 < \sigma, c_a \in [\frac{1}{M}, M] \right\}. \tag{44}$$

□

Proof of Lemma 4.4 We first prove that the problem equation 18 fulfills the linear independence constraint qualifications. For all $i = m+1, \ldots, n+1$, we denote the constraints of problem equation 18 as

$$h_i(u, v, c_a, \sigma) := S_i - u_i + v_i - \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp(-\frac{(a_i - a^*)^2}{2\sigma})}{\hat{\pi}(a_i, x_i)}. \tag{45}$$

The gradient of h_i is given by

$$\nabla h_i(u, v, c_a, \sigma) = \begin{bmatrix} \frac{\partial h_i}{\partial u_{m+1}}(u, v, c_a, \sigma) \\ \frac{\partial h_i}{\partial v_{m+1}}(u, v, c_a, \sigma) \\ \vdots \\ \frac{\partial h_i}{\partial u_i}(u, v, c_a, \sigma) \\ \frac{\partial h_i}{\partial v_i}(u, v, c_a, \sigma) \\ \vdots \\ \frac{\partial h_i}{\partial c_a}(u, v, c_a, \sigma) \\ \frac{\partial h_i}{\partial \sigma}(u, v, c_a, \sigma) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ -\frac{1}{\sqrt{2\pi}\sigma} \frac{\exp(-\frac{(a_i - a^*)^2}{2\sigma})}{\hat{\pi}(a_i, x_i)} \\ \frac{\exp(-\frac{(a_i - a^*)^2}{2\sigma})}{\hat{\pi}(a_i, x_i)} (\frac{c_a}{\sqrt{2\pi\sigma^2}} - \frac{c_a (a_i - a^*)^2}{2\sqrt{2\pi}\sigma^4}) \end{bmatrix}. \tag{46}$$

Therefore, with $\nabla h := (\nabla h_{m+1}, \ldots, \nabla h_{n+1})$ and $\lambda \in \mathbb{R}^{n+1}$, we obtain

$$\nabla h \cdot \lambda = 0 \iff \lambda = 0 \in \mathbb{R}^{n+1}. \tag{47}$$

As a result, the constraints are linearly independent. This property suffices for the KKT conditions to hold at any (local) optimum of equation 18.To furthermore show that the KKT conditions are also sufficient for a global optimum, we show that equation 18 is Type-I invex. An optimization problem with objective function $f(x)$ and constraints $g(x) <= 0$ with $x \in \mathbb{R}^{n+1}$ is Type-I invex at x_0, if there exists $\nu(x, x_0) \in \mathbb{R}^{n+1}$, such that

$$f(x) - f(x_0) \geq \nu(x, x_0)^T \nabla f(x_0), \tag{48}$$

and

$$-g(x_0) \geq \nu(x, x_0)^T \nabla g(x_0) \tag{49}$$

[20]. In problem equation 18, the gradients of the objective function and of each constraint h_i for all $i, j = m+1, \ldots, n+1$ at x_0 are given by

$$\frac{\partial \text{obj}(u_0, v_0, c_{a_0}, \sigma_0)}{\partial u_i} = 1 - \alpha, \quad \frac{\partial \text{obj}(u_0, v_0, c_{a_0}, \sigma_0)}{\partial v_i} = \alpha, \tag{50}$$

$$\frac{\partial \text{obj}(u_0, v_0, c_{a_0}, \sigma_0)}{\partial c_a} = \frac{\partial \text{obj}(u_0, v_0, c_{a_0}, \sigma_0)}{\partial \sigma} = 0 \tag{51}$$

$\forall i = m+1, \ldots, n+1$ and

$$\frac{\partial h_i}{\partial u_j}|_{u_0, v_0, c_{a_0}, \sigma_0} = \begin{cases} -1, & \text{for } i = j, \\ 0, & \text{else,} \end{cases} \qquad \frac{\partial h_i}{\partial v_j}|_{u_0, v_0, c_{a_0}, \sigma_0} = \begin{cases} 1, & \text{for } i = j, \\ 0, & \text{else,} \end{cases} \tag{52}$$

$$\frac{\partial h_i}{\partial c_a}|_{u_0, v_0, c_{a_0}, \sigma_0} = -\frac{1}{\sqrt{2\pi}\sigma_0} \frac{\exp\left(-\frac{(a_i - a^*)^2}{2\sigma_0}\right)}{\hat{\pi}(a_i, x_i)}, \tag{53}$$

$$\frac{\partial h_i}{\partial \sigma}|_{u_0, v_0, c_{a_0}, \sigma_0} = \frac{\exp\left(-\frac{(a_i - a^*)^2}{2\sigma_0}\right)}{\hat{\pi}(a_i, x_i)} \left(\frac{c_{a_0}}{\sqrt{2\pi}\sigma_0^2} - \frac{c_{a_0}(a_i - a^*)^2}{2\sqrt{2\pi}\sigma_0^4} \right). \tag{54}$$

For

$$\eta((u, v, c_a, \sigma), (u_0, v_0, c_{a_0}, \sigma_0)) := (-u_{0_1}, \ldots, -u_{0_{n+1}}, -v_{0_1}, \ldots, -v_{0_{n+1}}, -c_{a_0}, 0)^T, \tag{55}$$

the definition of Type-I invexity holds for equation 18. Thus, the KKT conditions are also sufficient for a global optimum. □

D.2 Proof of Theorem 4.2

We prove Theorem 4.2 in three steps: (i) We show that function class $\mathcal{F} := \{\theta \frac{\pi(a+\Delta_A|x)}{\pi(a|x)} \mid \theta \in \mathbb{R}^+\}$ indeed satisfies Eq. equation 2 for the intervention $a^* = a + \Delta_A$ and rewrite Eq. equation 5 as a convex optimization problem. (ii) We retrieve the corresponding dual problem, derive a dual prediction set, and show the equality of the coverage guarantee of the dual and the primal prediction sets. (iii) We derive S^* from the dual prediction set to construct C_{n+1} and prove the overall coverage guarantee. For further theoretical background on the idea of the proof, we refer to Gibbs et al. [18].

Justification of the distribution shift Observe that $\mathbb{E}[f(A, X)] = \theta$ for all $f \in \mathcal{F} := \{\theta \frac{\pi(a+\Delta_A|x)}{\pi(a|x)} \mid \theta \in \mathbb{R}^+\}$. Therefore, Eq. equation 2 simplifies to

$$\tilde{\pi}(a, x) = \frac{\pi(a + \Delta_A \mid x)}{\pi(a \mid x)} \pi(a \mid x) = \pi(a + \Delta_A \mid x). \tag{56}$$

Thus, $\mathcal{F}$ satisfies the propensity shift from Eq. equation 2 for the soft intervention $a^* = a + \Delta_A$. Following Lemma 4.1, we thus aim to find

$$\hat{q}_S := \arg\min_{\theta > 0} \frac{1}{n - m} \left(\sum_{i=m+1}^{n} l_\alpha(\theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)}, S_i) + l_\alpha(\theta \frac{\pi(a^* \mid x_{n+1})}{\pi(a_{n+1} \mid x_{n+1})}, S) \right). \tag{57}$$

Dual problem formulation First, we rewrite the primal problem as

$$\begin{aligned} \min_{\theta>0} \quad & \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i \\ \text{s.t.} \quad & S_i - \theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)}) - u_i + v_i = 0, \quad \forall i = m+1 \ldots, n+1, S_{n+1} = S \\ & u_i, v_i \geq 0, \quad \forall i = m+1 \ldots, n+1. \end{aligned} \tag{58}$$

For a reference, see [18]. The Lagrangian of the primal problem states

$$\mathcal{L} = \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i + \sum_{i=m+1}^{n+1} \eta_i \left(S_i - \theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)} - u_i + v_i \right) - \sum_{i=m+1}^{n+1} (\gamma_{1_i} u_i + \gamma_{2_i} v_i). \tag{59}$$

Setting derivative of $\mathcal{L}$ w.r.t. u_i and v_i to 0 results in

$$\frac{\partial \mathcal{L}}{\partial u_i} = (1-\alpha) - \eta_i - \gamma_{1_i} \stackrel{!}{=} 0, \quad \forall i = m+1 \ldots, n+1 \tag{60}$$

$$\frac{\partial \mathcal{L}}{\partial v_i} = (1-\alpha) - \eta_i - \gamma_{2_i} \stackrel{!}{=} 0, \quad \forall i = m+1 \ldots, n+1. \tag{61}$$

Since $\gamma_{1_i}, \gamma_{2_i} \geq 0 \; \forall i$, it follows for all $i = m+1, \ldots, n+1$ that

$$(1-\alpha) - \eta_i \geq 0 \quad \text{and} \quad \alpha - \eta_i \geq 0 \quad \Rightarrow -\alpha \leq \eta_i \leq 1-\alpha. \tag{62}$$

Therefore, the dual problem is formulated as

$$\begin{aligned} \max_{\eta_i, i=m+1,\ldots,n+1} \min_{\theta>0} \quad & \sum_{i=m+1}^{n} \eta_i \left(S_i - \theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)} \right) + \eta_{n+1} \left(S - \theta \frac{\pi(a^* \mid x_{n+1})}{\pi(a_{n+1} \mid x_{n+1})} \right) \\ \text{s.t.} \quad & -\alpha \leq \eta_i \leq 1-\alpha, \quad \forall i = m+1, \ldots, n+1. \end{aligned} \tag{63}$$

Coverage guarantee Recall from Lemma 4.1 that, for

$$\hat{q}_{S_{n+1}}(y) = \arg\min_{\theta>0} \frac{1}{n-m} \Big(\sum_{i=m+1}^{n} l_\alpha(\theta \frac{\pi(a_i + \Delta_A \mid x_i)}{\pi(a_i \mid x_i)}, S_i) + l_\alpha(\theta \frac{\pi(a^* \mid x_{n+1})}{\pi(a_{n+1} \mid x_{n+1})}, S_{n+1}(y))\Big), \tag{64}$$

we can construct $C_{n+1} = \{y \mid S_{n+1}(y) \leq \hat{q}_{S_{n+1}}(y)\}$ to achieve the desired coverage guarantee

$$P_f(Y(a^*) \in C(X_{n+1}, A^*(X_{n+1}))) \geq 1-\alpha. \tag{65}$$

It is infeasible to calculate $\hat{q}_{S_{n+1}}(y)$ directly. Therefore, we optimize the dual problem to receive

$$C(X_{n+1}, a^*) := \{y \mid S_{n+1}(y) \leq S^*\} \tag{66}$$

with S^* the maximum S, s.t. for η_{n+1}^S maximizing the dual problem, $\eta_{n+1}^S < 1-\alpha$. Hence, it is left to show that replacing $\hat{q}_{S_{n+1}}(y)$ by S^* in C does not change to coverage guarantee.

To do so, we fix some $\theta > 0$ to obtain a specific $f(a,x) := \theta \frac{\pi(a^*|x)}{\pi(a|x)}$. Let $\hat{g}(a,x) \in \mathcal{F}$ denote the primal optimal solution. Recall the Lagrangian

$$\mathcal{L} = \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i + \sum_{i=m+1}^{n+1} \eta_i (S_i - f(a_i, x_i) - u_i + v_i) - \sum_{i=m+1}^{n+1} (\gamma_{1_i} u_i + \gamma_{2_i} v_i). \tag{67}$$

Deriving wrt. f yields the stationarity condition

$$0 \stackrel{!}{=} - \sum_{i=m+1}^{n+1} \eta_i^S f(a_i, x_i) \tag{68}$$

$$= - \sum_{S_i < \hat{g}(a_i,x_i)} \eta_i^S f(a_i, x_i) - \sum_{S_i > \hat{g}(a_i,x_i)} \eta_i^S f(a_i, x_i) - \sum_{S_i = \hat{g}(a_i,x_i)} \eta_i^S f(a_i, x_i). \tag{69}$$

The complementary slackness Karush-Kuhn-Tucker conditions yield

$$\eta_i^S \in \begin{cases} -\alpha, & \text{if } S_i < \hat{g}(a_i, x_i), \\ [-\alpha, 1-\alpha], & \text{if } S_i = \hat{g}(a_i, x_i), \\ 1-\alpha, & \text{if } S_i > \hat{g}(a_i, x_i). \end{cases} \tag{70}$$

Therefore, we can rewrite the equation from above as

$$0 = \sum_{S_i < \hat{g}(a_i,x_i)} \alpha f(a_i, x_i) - \sum_{S_i > \hat{g}(a_i,x_i)} (1-\alpha) f(a_i, x_i) - \sum_{S_i = \hat{g}(a_i,x_i)} \eta_i^S f(a_i, x_i) \tag{71}$$

$$= \sum_{\eta_i^S < 1-\alpha} \alpha f(a_i, x_i) - \sum_{\eta_i^S = 1-\alpha} (1-\alpha) f(a_i, x_i) - \sum_{\substack{\eta_i^S < 1-\alpha, \\ S_i = \hat{g}(a_i,x_i)}} (\alpha + \eta_i^S) f(a_i, x_i) \tag{72}$$

$$= \sum_{i=m+1}^{n+1} (\alpha - \mathbb{1}_{[\eta_i^S = 1-\alpha]}) f(a_i, x_i) - \sum_{\substack{\eta_i^S < 1-\alpha, \\ S_i = \hat{g}(a_i,x_i)}} (\alpha + \eta_i^S) f(a_i, x_i). \tag{73}$$

Before deriving the coverage guarantee from the stationarity condition, we state the following lemma to underline the definition of S^*.

Lemma D.1 (Gibbs et al. [18]). *The mapping $S \mapsto \eta_{n+1}^S$ is non-decreasing in S for all η_{n+1}^S maximizing*

$$\begin{aligned} \max_{\eta_i, i=m+1,\ldots,n+1} \min_{g \in \mathcal{F}} \quad & \sum_{i=1}^{n} \eta_i (S_i - g(a_i, x_i)) + \eta_{n+1}(S - g(a_{n+1}, x_{n+1})) \\ \textit{s.t.} \quad & -\alpha \leq \eta_i \leq 1-\alpha, \quad \forall i = m+1, \ldots, n+1 \end{aligned} \tag{74}$$

for non-negative function classes $\mathcal{F}$.

To prove the final coverage guarantee, we observe that

$$\mathbb{E}[f(a_{n+1}, x_{n+1})(\mathbb{1}_{[Y(a^*) \in C(X_{n+1}, a^*)]} - (1-\alpha))] \tag{75}$$

$$= \mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[Y(a^*) \notin C(X_{n+1}, a^*)]})] \tag{76}$$

$$= \mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[S(y) > S^*]})]. \tag{77}$$

With the definition of S^* as the maximum optimizer η_{n+1}^S with $\eta_{n+1}^S < 1-\alpha$ and Lemma D.1, it follows that

$$\mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[S(y) > S^*]})] = \mathbb{E}[(\alpha - \mathbb{1}_{[\eta_{n+1}^S = 1-\alpha]}) f(a_{n+1}, x_{n+1})] \tag{78}$$

and, by exchangeability of $(f(a_i, x_i), \hat{q}_S(a_i.x_i), S_i)$, that

$$\mathbb{E}[(\alpha - \mathbb{1}_{[\eta_i^S = 1-\alpha]}) f(a_i, x_i)] = \mathbb{E}\left[\frac{1}{n-m} \sum_{i=m+1}^{n+1} (\alpha - \mathbb{1}_{[\eta_i^S = 1-\alpha]}) f(a_i, x_i)\right] \tag{79}$$

$$= \mathbb{E}\left[\frac{1}{n-m} \sum_{\substack{\eta_i^S < 1-\alpha, \\ S_i = \hat{g}(a_i,x_i)}} (\alpha + \eta_i^S) f(a_i, x_i)\right]. \tag{80}$$

Since f is positive and $\eta_i \in [-\alpha, 1-\alpha]$, it follows

$$\mathbb{E}[f(a_{n+1}, x_{n+1})(\mathbb{1}_{[Y(a^*) \in C(X_{n+1}, a^*)]} - (1-\alpha))] \geq 0 \tag{81}$$

and thus

$$P_f(Y(a^*) \in C_{n+1} C(X_{n+1}, A^*(X_{n+1}))) \geq 1-\alpha. \tag{82}$$

□

D.3 Proof of Theorem 4.5

We follow the same outline as in the proof of Theorem 4.2 in Section D.2. In Lemma 4.3, we motivated the functional class of distribution shifts. Therefore, it is left to prove the coverage guarantee of C_{n+1}.

Key to our proof is the following lemma.

Lemma D.2. *The mapping $S \mapsto v_{n+1}^S$ is non-increasing in S for all $g^S(x,a)$ minimizing*

$$\begin{aligned} \min_{g \in \mathcal{F}} \quad & \sum_{i=m+1}^{n+1} (1-\alpha)u_i + \alpha v_i \\ \text{s.t.} \quad & S_i - g(x_i, a_i) - u_i + v_i = 0, \quad \forall i = m+1, \ldots, n+1 \end{aligned} \tag{83}$$

for non-negative function classes $\mathcal{F}$ and imputed $S_{n+1} = S$ stemming from a non-negative non-conformity score function (e.g., the residual of the prediction).

Proof. Assume for contradiction that there exists $\tilde{S} > S$ such that $v_{n+1}^{\tilde{S}} > v_{n+1}^S$. Then

$$(\tilde{S} - S)(v_{n+1}^{\tilde{S}} - v_{n+1}^S) > 0. \tag{84}$$

We observe that

$$\tilde{S}(S - g^S(x_{n+1}, a_{n+1}) - u_{n+1}^S + v_{n+1}^S) = S(\tilde{S} - g^{\tilde{S}}(x_{n+1}, a_{n+1}) - u_{n+1}^{\tilde{S}} + v_{n+1}^{\tilde{S}}) = 0. \tag{85}$$

Reformulating the equation above yields

$$\begin{aligned} & (\tilde{S} - S)(v_{n+1}^{\tilde{S}} - v_{n+1}^S) && (86)\\ = \;& \tilde{S}u_{n+1}^S - Su_{n+1}^{\tilde{S}} + \tilde{S}g^S(x_{n+1}, a_{n+1}) - Sg^{\tilde{S}}(x_{n+1}, a_{n+1}) + \tilde{S}v_{n+1}^{\tilde{S}} - Sv_{n+1}^S && (87)\\ < \;& S(u_{n+1}^S - u_{n+1}^{\tilde{S}} + g^S(x_{n+1}, a_{n+1}) - g^{\tilde{S}}(x_{n+1}, a_{n+1}) - (v_{n+1}^S - v_{n+1}^{\tilde{S}})) && (88)\\ = \;& S(S - \tilde{S}). && (89) \end{aligned}$$

This is equivalent to

$$\begin{aligned} & (S - \tilde{S})(v_{n+1}^S - v_{n+1}^{\tilde{S}}) < S(S - \tilde{S}) && (90)\\ \iff \quad & v_{n+1}^S - v_{n+1}^{\tilde{S}} > S \geq 0, && (91) \end{aligned}$$

which contradicts the assumption that $v_{n+1}^{\tilde{S}} > v_{n+1}^S$. □

Coverage guarantees. As in D.2, we fix some $\sigma > 0$ and $c_a \in [\frac{1}{M}, M]$ to obtain a specific $f(a,x) := \frac{c_a}{\sqrt{2\pi}\sigma} \frac{\exp\left(-\frac{(a_i - a^*)^2}{2\sigma^2}\right)}{\hat{\pi}(a_i|x_i)}$. We further denote $\hat{g}(a,x) \in \mathcal{F}$ the optimal solution given by the optimal values $\hat{\sigma}$ and $\hat{c_a}$.

With the definition of S^* as the minimum S such that $v_{n+1}^{S^*} = 0$ and Lemma D.2, we now can state

$$\begin{aligned} \mathbb{E}[f(a_{n+1}, x_{n+1})(\mathbb{1}_{[Y(a^*) \in C_{n+1}]} - (1-\alpha))] &= \mathbb{E}[f(a_{n+1}, x_{n+1})(\mathbb{1}_{[v_{n+1}^S > 0]} - (1-\alpha))] && (92)\\ &= \mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[v_{n+1}^S = 0]})] && (93) \end{aligned}$$

and, by exchangeability of $(f(a_i, x_i), \hat{q}_S(a_i.x_i), S_i)$, that

$$\begin{aligned} \mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[v_{n+1}^S = 0]})] &= \mathbb{E}\left[\frac{1}{n-m} \sum_{i=m+1}^{n+1} f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[v_{n+1}^S = 0]})\right] && (94)\\ &= \frac{1}{n-m}\mathbb{E}\left[\sum_{v_i^S > 0} \alpha f(a_i, x_i) - \sum_{v_i^S = 0})(1-\alpha) f(a_i, x_i)\right] && (95)\\ &= \frac{1}{n-m}\mathbb{E}\left[\sum_{S_i < \hat{g}(a_i, x_i)} \alpha f(a_i, x_i) - \sum_{S_i \geq \hat{g}(a_i, x_i)})(1-\alpha) f(a_i, x_i)\right]. && (96) \end{aligned}$$

Deriving the Lagrangian above wrt. f yields the stationarity condition

$$0 \stackrel{!}{=} \sum_{i=m+1}^{n+1} \eta_i^S f(a_i, x_i) \tag{97}$$

$$= \sum_{S_i < \hat{g}(a_i, x_i)} \eta_i^S f(a_i, x_i) + \sum_{S_i > \hat{g}(a_i, x_i)} \eta_i^S f(a_i, x_i) + \sum_{S_i = \hat{g}(a_i, x_i)} \eta_i^S f(a_i, x_i). \tag{98}$$

The complementary slackness Karush-Kuhn-Tucker conditions yield

$$\eta_i^S \in \begin{cases} -\alpha, & \text{if } S_i < \hat{g}(a_i, x_i), \\ [-\alpha, 1-\alpha], & \text{if } S_i = \hat{g}(a_i, x_i), \\ 1-\alpha, & \text{if } S_i > \hat{g}(a_i, x_i). \end{cases} \tag{99}$$

Therefore, we receive

$$\mathbb{E}[f(a_{n+1}, x_{n+1})(\alpha - \mathbb{1}_{[v_{n+1}^S = 0]})] = \frac{1}{n-m} \mathbb{E}\left[\sum_{\eta_i^S < 1-\alpha} \alpha f(a_i, x_i) - \sum_{\eta_i^S = 1-\alpha})(1-\alpha) f(a_i, x_i) \right] \tag{100}$$

$$= \frac{1}{n-m} \mathbb{E}\left[\sum_{\substack{\eta_i^S < 1-\alpha, \\ S_i = \hat{g}(a_i, x_i)}} (\alpha + \eta_i^S) f(a_i, x_i) \right]. \tag{101}$$

Since f is positive and $\eta_i \in [-\alpha, 1-\alpha]$, it follows

$$\mathbb{E}[f(a_{n+1}, x_{n+1})(\mathbb{1}_{[Y(a^*) \in C(X_{n+1}, a^*)]} - (1-\alpha))] \geq 0 \tag{102}$$

and thus

$$P_f(Y(a^*) \in C(X_{n+1}, a^*)) \geq 1 - \alpha. \tag{103}$$

□

E Additional background

E.1 Extended literature review

Uncertainty quantification for causal quantities

There exist various methods for uncertainty quantification of causal quantities. These are often based on Bayesian methods [e.g., 1, 21, 22, 24]. However, Bayesian methods require the specification of a prior distribution based on domain knowledge and are thus neither robust to model misspecification nor generalizable to model-agnostic machine learning models. Other methods only provide asymptotic guarantees [e.g., 25, 27]. The strength of conformal prediction, however, is to provide finite-sample uncertainty guarantees.

In the following, we present related work on CP for causal quantities in more detail.

Recently, Alaa et al. [2] provided predictive intervals for CATE meta-learners under the assumption of full knowledge of the propensity score. As an extension, Jonkers et al. [27] proposed a Monte-Carlo sampling approach to receive less conservative intervals. Chen et al. [9] provide prediction intervals for counterfactual outcomes. However, the proposed method requires access to additional interventional data and is thus not applicable to real-world applications on observational data. All methods are restricted to binary treatments.

Other works focus on prediction intervals for off-policy prediction [47, 58] and conformal sensitivity analysis [56], thus neglecting estimation errors arising from propensity or weight estimation or for randomized control trials [30]. Wang et al. [52] constructed intervals with treatment-conditional coverage of discrete treatments. Aiming for group-conditional coverage, Wang et al. [52] adapted CP to cluster randomized trials. Nevertheless, the method only applies to a finite number of treatments and thus is not applicable to continuous treatments.

Lei and Candès [36] consider the estimated propensity by incorporating the estimation error as a TV-distance term in the coverage guarantees. However, for large TV-distances (close to 1), the proposed method can only construct intervals with a very limited coverage $\alpha \in (0, 1 - TV)$. Hence, the method is not suitable for applications in medical practice. Our method, however, can also construct intervals with high coverage guarantees for high estimation errors. An increased error will widen the prediction intervals instead of reducing the coverage guarantee. We consider our approach more suitable for medical practice, as one can visually inspect the intervals and decide on the suitability of the task at hand.

Overall, no method can provide exact intervals for continuous treatments. Especially, no method considers the error arising from propensity estimation in the analysis.

Conformal prediction under covariate shift Multiple works on CP with *marginal coverage* under distribution shifts between training and test data have been introduced in the literature [e.g., 8, 11, 15, 16, 17, 18, 19, 36, 41, 48, 55]. Our setting also involves a distribution shift due to the intervention on the treatment but differs from the latter in that the true distribution shift is unknown.

Gibbs et al. [18] introduced an approach to derive CP intervals under unknown distribution shifts. It proves valid finite-sample prediction intervals for all distribution shifts in a finite-dimensional function class. However, the approach does **not** directly apply to causal inference settings. Nevertheless, our framework builds upon the work by Gibbs et al. [18] in that we re-frame the proposed approach to apply to the distribution shift induced through the intervention in causal effect estimation. In this setting, the distribution shift is captured by the shift of the propensity function. Adapting Gibbs et al. [18] to a causal inference setting requires carefully addressing the underlying challenges that come from computing CP intervals in a causal inference setting (e.g., propensity score estimation, hard/soft interventions), which we regard as our main novelty and which is of immediate practical relevance (e.g., in personalized medicine).

E.2 The need for exchangeability in CP

Coverage guarantees of existing CP intervals essentially rely on the exchangeability of the nonconformity scores. Exchangeability assures that the nonconformity score of the test point $n + 1$ is equally likely to fall anywhere among the calibration scores, its rank is uniform, and that uniformity

is exactly what yields the distribution-free coverage guarantee. Without exchangeability, the rank is not guaranteed to be uniform, and the marginal coverage bound can fail.

However, intervening on treatment A shifts the propensity function and, therefore, induces a shift in the covariates between calibration and test data, specifically in treatment A. Therefore, exchangeability is not fulfilled, and the coverage guarantees might fail.

As a remedy, we present a novel and powerful remedy in our work: The overall distribution of the confounders X is assumed to stay constant between train, calibration, and test data (as standard in ML problems). This is completely orthogonal to constructing intervals for different (e.g., young or old) patients. Note that CP intervals are constructed for only one sample/patient at a time. This means that different intervals are constructed for patients with different features X. In other words, the intervals are constructed conditionally on X, but the coverage guarantee is marginal across the complete population of X. Overall, the shift from one patient to another does not pose any challenges for CP methods.

F Extended discussion

F.1 Discussion on the tightness of our CP intervals:

Our method builds upon the idea of CP to provide finite-sample coverage guarantees. Notably, standard CP does not provide intervals that are proven to be sharp. To our knowledge, there is no method that provides sharp/the tightest possible intervals for potential outcomes. Exploring the tightness of our CP intervals is an interesting and important direction for future research.

In practice, it is possible to observe the width, and thus the informativeness for decision-making, of the intervals. However, coverage guarantees cannot be observed. Therefore, we help the decision-maker by providing valid intervals. The decision-maker has to decide, case by case, if the returned intervals are beneficial for the problem at hand. Note that this aspect of the informativeness of intervals holds true for any uncertainty quantification method (including those to be proven to be sharp).

F.2 A note on challenges and difficulties in CP for causal effects of continuous treatments

Existing works on conformal prediction for binary or (low-dimensional) discrete treatment are commonly based on (a) weighted conformal prediction [48] or (b) conformal prediction local coverage guarantees [35]. The first approach provides marginal coverage under a distribution shift through reweighting. It requires computing the weights based on the probability of treatment $A = a$. However, for continuous treatments, this is always zero. Although applicable to binary or low-dimensional discrete treatments [e.g., 36], this weighting approach cannot be extended similarly to continuous treatments. Furthermore, the propensity of a continuous treatment given by the Dirac delta function δ_a would require us to restrict the calibration to data samples of the specific treatment, which are extremely rare or even *might be missing*. Therefore, the calibration step cannot be employed in our setting. The second approach provides treatment group-conditional coverage. Although again possible for binary or low-dimensional treatments, this approach *does not apply to continuous treatment*s as no treatment groups can be defined. Instead, we propose a novel method for conformal predictions that circumvents the above problems and is carefully tailored to continuous treatments.

F.3 Causal effects of continuous treatments & kernel smoothing

Causal inference becomes challenging with continuous treatments primarily due to the infinite number of potential outcomes per sample, from which only one outcome is observed. Continuous treatments thus result in causal effects that are generally represented by curves (called dose-response curves) [29]. This is unlike binary treatment, where the causal effects are represented by a single discrete value.

For continuous treatments, the dose-response curves are typically assumed to fulfill some smoothness criterion [e.g., 40, 45]. Hence, when estimating treatment effects, interpolation and kernel smoothing of the outcome function are commonly employed [e.g., 29, 37].

Underlying causal estimation with continuous treatments is the generalized propensity score [23]. It is defined as the conditional probability of receiving treatment a^* given the covariates X under the following regularity conditions: (i) For each i, $Y_i(a), x_i, A_i$ are defined on a common probability space; (ii) A_i is continuously distributed with respect to the Lebesgue measure; and (iii) $Y_i = Y_i(A_i)$ is a well-defined random variable.

Approximating the density $\delta_{a^*}(a)$ of the hard intervention a^* through a Gaussian kernel follows directly from the definition of $\delta_{a^*}(a)$ as the limit of such kernel. This is also common in the literature [e.g., 28]. Importantly, we note that we do not directly approximate the potential outcome $Y(a^*)$ (but only the propensity scores). Thus, we do not have a bias-variance trade-off of the estimated outcome. Due to the smoothness of the dose-response curve, it is now valid to employ observed samples within a treatment region of a^* defined by σ to construct the intervals. We note that the importance of the samples is weighted by the inverse distance of the sample to a^* in treatment space. We give further intuition on the relationship between σ, the importance of observational samples, and the prediction interval width in the following.

F.4 Interpretation of optimal parameters

To obtain CP intervals under an unknown distribution shift, we approximate the Dirac-delta distribution representing the hard intervention by a Gaussian function as

$$\delta_{a^*}(a) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-a^*)^2}{2\sigma^2}\right). \tag{104}$$

In Theorem 4.5, we thus optimize over $\sigma > 0$ and $c_a \in [\frac{1}{M}, M]$ to obtain the $(1-\alpha)$-quantile of the distribution shift-calibrated non-conformity scores. The optimal parameter σ^* represents a trade-off between the uncertainty in the prediction and the uncertainty in the interval construction: A small σ^* resembles the propensity of the hard intervention best. Thus, with sufficient or even infinite data close to a^*, we could construct the narrowest CP interval. However, the smaller σ, the less data close to a^* will be available to calculate the prediction interval in practice. As a result, many calibration data samples will be strongly perturbed during the calculation, which increases the uncertainty and, thus, the interval size.

The parameter c_a allows us to incorporate the estimation error in the propensity score. It represents a weighting of the propensity shift such that the $(1-\alpha)$-quantile of the non-conformity scores is increased with higher estimation error.

F.5 Interpretation of parameter M

Our optimization requires the specification of a parameter M, denoting a bound on the propensity estimation error. One can view the parameter M as a type of sensitivity parameter. Therefore, we follow former work in causal inference and propose to incorporate domain knowledge to specify the parameter M [e.g., 13**?**]. Another way of making use of the parameter M is to observe how the intervals change for varying M. This indicates how much effect the propensity misspecification has on the prediction interval and can help in making reliable decisions. A third option to calibrate M is to employ measures for epistemic uncertainty on top of the propensity estimate when there is no domain knowledge for specifying M.

F.6 A note on the stability of our method

In our experiments, one can observe some instability for certain privacy budgets and intervention combinations. This is likely due to the fact that the CP coverage guarantees are only *marginal*. Therefore, we might experience under- or over-coverage. However, we note that across all runs, our method, on average, achieves the desired coverage. These instabilities only occur in single settings. Furthermore, the variance in coverage of our method is much lower than the coverage variance of the baselines. A more in-depth analysis of the stability of the proposed method is left for future work.

G Experimentation details

G.1 Synthetic dataset generation

We consider two different propensity and outcome functions. In each setting, we assign two types of interventions: a known propensity shift of $\Delta = 1, 5, 10$, i.e., three soft interventions $a^* = a + \Delta$, and the point interventions $a^* \in \{1x, 5x, 10, \}$ given the confounder $X = x$.

We generate synthetic datasets for each setting. Specifically, we draw each 2000 train, 1000 calibration, and each 1000 test samples per intervention from the following structural equations. Dataset 1 is given by

$$
\begin{aligned}
&X \sim \text{Uniform}[1, 4] \quad \text{(integer)} \\
&A \sim p \cdot \text{Uniform}[0, 5X) + (1 - p)\text{Uniform}[5X, 40], \quad p \sim \text{Bernoulli}(0.3) \\
&Y \sim \sin\left(\frac{\pi}{6}(0.1A - 0.5X)\right) + \text{Normal}(0, 0.1),
\end{aligned}
$$

and dataset 2 by

$$
\begin{aligned}
&X \sim \text{Uniform}[1, 4] \quad \text{(integer)} \\
&A \sim \text{Normal}(5X, 10) \\
&Y \sim \sin\left(\frac{\pi}{2}(0.1A - 0.1X)\right) + \text{Normal}(0, 0.1),
\end{aligned}
$$

G.2 Medical dataset

We use the MIMIC-III dataset [26], which includes electronic health records (EHRs) from patients admitted to intensive care units. From this dataset, we extract 8 confounders (heart rate, sodium, blood pressure, glucose, hematocrit, respiratory rate, age, gender) and a continuous treatment (mechanical ventilation) using an open-source preprocessing pipeline [53]. From each patient trajectory in the EHRs, we sample random time points and average the value of each variable over the ten hours before the sampled time point. We define the variable blood pressure after treatment as the outcome, for which we additionally apply a transformation to be more dependent on the treatment and less on the blood pressure before treatment. We remove all patients (samples) with missing values and outliers from the dataset. Outliers are defined as samples with values smaller than the 0.1th percentile or larger than the 99.9th percentile of the corresponding variable. The final dataset contains 14719 samples, which we split into train (60%), val (10%), calibration (20%), and test (10%) sets.

G.3 Implementation details

Our experiments are implemented in PyTorch Lightning. We provide our code in our GitHub repository. All experiments were run on an AMD Ryzen 7 PRO 6850U 2.70 GHz CPU with eight cores and 32GB RAM.

We limited the experiments to standard multi-layer perception (MLP) regression models, consisting of three layers of width 16 with ReLu activation function and MC dropout at a rate of 0.1, optimized via Adam. We did not perform hyperparameter optimization, as our method aimed to provide an agnostic prediction interval applicable to any prediction model. All models were trained for 300 epochs with batch size 32.

Our algorithm requires solving (non-convex) optimization problems through mathematical optimization. We chose to employ two interior-point solvers in our experiments: For the experiments with soft interventions that pose convex optimization problems, we use the solver MOSEK. For the hard interventions, which included non-convex problems, we used the solver IPOPT. Both solvers were run with default parameters.

G.4 Selection of the interventions in our experiments

The treatment (in the complete dataset) is modeled to lie in the range [0,40]. Therefore, the treatments/interventions can also only fall into this range. All samples that would have achieved a treatment outside this range through the interventions were neglected in our analysis. To guarantee

that a sufficient number of samples were included in our experiments, we chose the maximal soft treatment as an increase of 10.

We sampled the covariates X in the range from 1 to 4. To again perform interventions that fall inside the range of [0,40], we decided to set the intervention as $7X$ and $10X$. Other choices of interventions would also have been possible. Reassuringly, there was no systematic selection of interventions in our experiments besides the considerations above.

For the soft interventions, we do not see different effects of the interventions on the two datasets. For the hard interventions, however, we can observe a difference. Recall that in dataset 1, the treatment was sampled uniformly from [0,40] (with a dependence on X), whereas in dataset 2, it was sampled from a normal distribution with a mean of $5X$. Therefore, the intervention $10X$ is far in the tail of the distribution. As a result, we observe a slightly lower coverage for this intervention on dataset 2 compared to dataset 1.

H Further results

We present further results from our experiments in Section 5. Specifically, we state the prediction performance of the underlying models ϕ, discuss the scalability of our approach, and show the prediction intervals per covariate for various significance levels α and soft interventions Δ of our synthetic experiments on dataset 1 and dataset 2.

Performance: We first report the performance of the underlying prediction models ϕ for the synthetic datasets across 50 runs in Table 4. The prediction model on the real-world dataset achieved a mean squared error loss of 1.2373.

	Dataset 1	Dataset 2	MIMIC
ϕ	0.0216 (0.0056)	0.9029 (0.3908)	0.0141 (0.0057)
Ens.	0.0094 (2.1169e^{-5})	0.0130 (0.0003)	-

Table 4: Mean and standard deviation of MSE loss of prediction models ϕ across 50 runs.

Width of the prediction intervals: We further report the width of the prediction intervals in our synthetic experiments in Table 5. The width is important to assess the usefulness of the resulting prediction intervals. As the performance of the ensemble method is not comparable with the coverage of MC-Dropout and our CP method, we only compare the latter two methods with regard to the interval width.

	Dataset 1		Dataset 2	
Delta	Ours	MC-Dropout	Ours	MC-Dropout
1	0.3647 (0.1284)	0.1938 (0.1170)	0.4051 (0.1036)	0.2897 (0.1480)
5	0.4024 (0.2285)	0.1653 (0.1103)	0.4610 (0.2479)	0.3036 (0.1455)
10	0.4301 (0.2610)	0.1639 (0.1080)	0.6711 (0.8520)	0.3235 (0.1445)

Table 5: Mean and standard deviation of the resulting prediction intervals.

Comparison to the vanilla CP baseline: We compare our method to the naive vanilla CP (V-CP), i.e., a CP method that does not account for the distribution shift. We observe that V-CP does not achieve any valid prediction interval across all distribution shifts and confidence levels. This can be explained by the good prediction performance of the underlying model. Thus, V-CP intervals are extremely small (average width of 0.0003) and can never cover the true potential outcome after the intervention. Overall, the results confirm the importance of accounting for the distribution shift induced by the intervention.

Scalability: Calculating the prediction intervals requires an iterative search for an optimal value S^*. Therefore, the underlying optimization problem must be fitted multiple times throughout the algorithm, potentially posing scalability problems. In our empirical studies, however, we did not encounter scalability issues. Importantly, we found that the average runtime of our algorithm on a standard desktop CPU is only 16.43 seconds. On the MIMIC dataset, computing CP intervals takes roughly 10 times longer than computing MC intervals. However, we emphasize that MC intervals are generally *not* faithful and therefore *not* directly comparable.

Prediction intervals: In Figures 8, 9, and 10, we present the prediction bands given by our

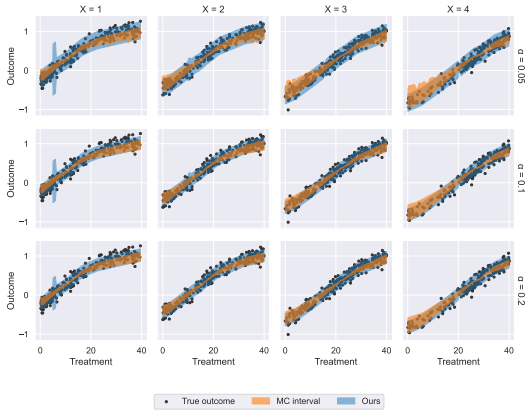

Figure 8: Prediction intervals for multiple significance levels α for the synthetic dataset 1 with intervention $\Delta = 1$.

method and MC dropout on dataset 1. In particular, for confounder $X = 1$, our method shows a large increase in the uncertainty in the potential outcomes of treatments affected by the intervention. MC dropout does not capture this uncertainty.

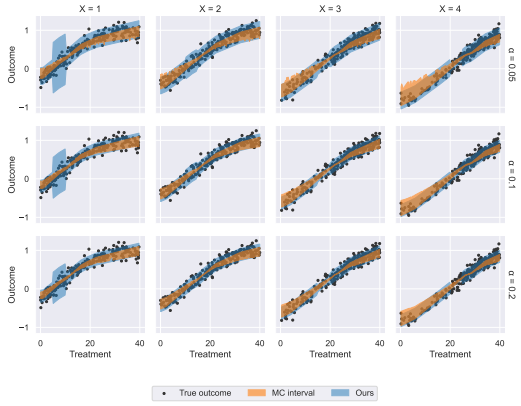

Figure 9: Prediction intervals for multiple significance levels α for the synthetic dataset 1 with intervention $\Delta = 5$

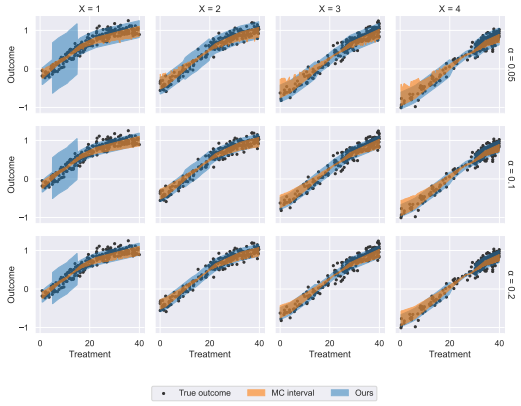

Figure 10: Prediction intervals for multiple significance levels α for the synthetic dataset 1 with intervention $\Delta = 10$.

In Figures 11 and 12, we present the prediction bands given by our method and MC dropout on dataset 2 for the soft interventions $\Delta = 1$ and $\Delta = 10$ (the results for $\Delta = 5$ were presented in the main paper). We observe that the prediction intervals for $\Delta = 10$ become extremely wide for high treatments. This aligns with our expectation, as data for high treatments in combination with low confounders is rare or even absent in the dataset. Thus, the expected uncertainty is very high.

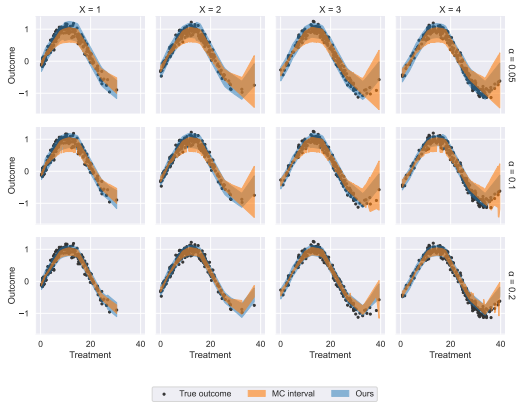

Figure 11: Prediction intervals for multiple significance levels α for the synthetic dataset 2 with intervention $\Delta = 1$

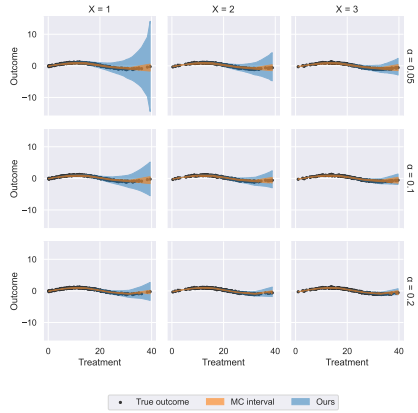

Figure 12: Prediction intervals for multiple significance levels α for the synthetic dataset 2 with intervention $\Delta = 10$.