Using representation balancing to learn conditional-average dose responses from clustered data

Anonymous authors
Paper under double-blind review

Abstract

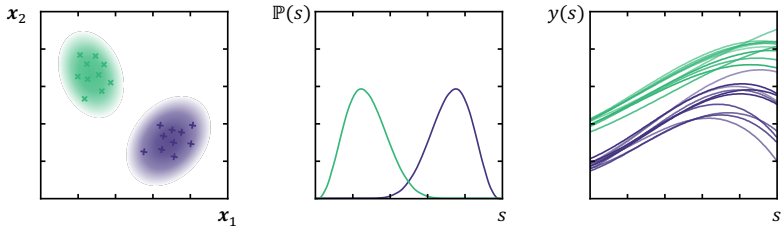
Estimating a unit's responses to interventions with an associated dose, the "conditional average dose response" (CADR), is relevant in a variety of domains, from healthcare to business, economics, and beyond. Such a response typically needs to be estimated from observational data, which introduces several challenges. That is why the machine learning (ML) community has proposed several tailored CADR estimators. Yet, the proposal of most of these methods requires strong assumptions on the distribution of data and the assignment of interventions, which go beyond the standard assumptions in causal inference. Whereas previous works have so far focused on smooth shifts in covariate distributions across doses, in this work, we will study estimating CADR from clustered data and where different doses are assigned to different segments of a population. On a novel benchmarking dataset, we show the impacts of clustered data on model performance and propose an estimator, CBR-Net, that learns cluster-agnostic and hence dose-agnostic covariate representations through representation balancing for unbiased CADR inference. We run extensive experiments to illustrate the workings of our method and compare it with the state of the art in ML for CADR estimation.

1 Introduction

Predicting conditional-average intervention responses has become a popular field of machine learning (ML) research (Shalit et al., 2016; Shi et al., 2019).[1] While a large share of studies focuses on estimating responses to categorical and especially binary-valued interventions (Shalit et al., 2016; Yoon et al., 2018; Shi et al., 2019; Johansson et al., 2020), less focus has been put on continuous-valued interventions, so interventions with an associated intensity or dose (Schwab et al., 2019). Yet, estimating the conditional average "dose response" (CADR) is of great interest in many domains of applications, such as marketing (Farrelly et al., 2005), medicine (Calabrese, 2016), or education (Turk, 2019), especially when too much or too little intervention can have negative effects (Frei & Canellos, 1980), or when decision-makers must optimize resource usage (Vanderschueren et al., 2023).

Next to high-dimensional covariates and non-linear dose-responses, an important challenge is for an estimator to adjust for confounding (Bica et al., 2020), so the presence of a covariate vector that influences the dose assignment, as well as the response (Haneuse, 2016). Not accounting for confounding might hinder methods from learning unbiased estimates of CADR by overfitting the dose assignment mechanism (Shalit et al., 2016). This is why the ML community has proposed several methods to learn CADR from observational data (Schwab et al., 2019; Bica et al., 2020; Nie et al., 2021; Wang et al., 2022). Next to the standard assumptions necessary for causal inference (Stone, 1993), most of these estimators rely on further assumptions on the underlying data, such as on the distribution of covariates and the assignment of interventions and doses (Nie et al., 2021; Wang et al., 2022). These assumptions drive algorithm development, yet are often not motivated by investigations of real-life scenarios and their accompanying data-generating processes (DGPs). This potentially explains the little use of ML estimators for intervention response estimation by practitioners (Curth et al., 2021). One such assumption is a smooth shift in covariate distributions across different doses,

[1] Some literature also refers to the conditional-average response as the "individual" response (Vegetabile, 2021).



(a) Input space (b) Dose distributions (c) Dose responses

Figure 1: **Illustration of confounding by cluster (CBC).** (a) We assume that there are clusters of similar units in the data. (b) The dose assignment mechanism is a function of the cluster so similar units are assigned similar doses. As dose responses are heterogeneous and depend on a unit's covariates (c), the resulting data is confounded.

as required by Wang et al. (2022) and satisfied in the synthetic benchmarking datasets by Bica et al. (2020) and Nie et al. (2021) used throughout the literature.

In this paper, we investigate CADR estimation from data in which units are clustered and where different doses are assigned to different segments of a population (cf. Figure 1). Such data is prevalent in many domains, such as loan and insurance pricing, where groups of similar customers receive similar prices (Phillips, 2013), or epidemiology, where the impacts of clustered data on modeling have been studied widely (Berlin et al., 1999; Seaman et al., 2014). In ML research, however, we find that little interest has been paid to understanding its impact on intervention and dose response estimates. In this paper, we take a step towards closing this gap in the literature. By analyzing the real-world applications of CADR estimation in healthcare, public policy, and business, we abstract DGPs for clustered data and create a novel semi-synthetic benchmarking dataset for dose response estimation from clustered data with varying levels of confounding. We evaluate several ML methods, including traditional supervised learning methods and tailored ML methods for CADR estimation. In addition, we propose a new method for estimating CADR from clustered observational data with cluster-based dose assignment. Our method is called CBRNet (**C**luster-robust dose response estimation through **B**alanced **R**epresentation learning with neural **Net**works, cf. Figure 3) and is motivated by preceding studies on estimating intervention responses under confounding through the balancing of covariate distributions (Shalit et al., 2016; Schwab et al., 2018).

Contributions. Our paper makes three important contributions to the ML literature on dose response estimation: (1) To the best of our knowledge, we are the first to investigate the impacts of learning from clustered data on the performance of ML estimators for CADR; (2) we propose a novel semi-synthetic benchmark that simulates CBC in observational data and that allows for testing its impacts on estimators; (3) we introduce CBRNet, a method for CADR estimation from clustered data, that leverages representation balancing. We perform extensive experiments to test the performance of CBRNet, both on our newly presented benchmark and on previously established datasets.

Outline. Section 2 discusses related work on ML methods for CADR estimation and preceding benchmarking practices. Our problem formulation follows in Section 3. We present our method, CBRNet, in Section 4, discussing motivation and architecture. Our experimental setup is presented in Section 5, with results following in Section 6. We conclude with Section 7, discussing the impact of our work and future research directions.

2 Context

Background. The majority of ML literature on estimating intervention responses has been concerned with "treatment effects", so the differences in outcome between applying a specific intervention, the "treatment", or not (Johansson et al., 2016; Shalit et al., 2016; Louizos et al., 2017; Nie & Wager, 2017; Wager & Athey, 2018; Shi et al., 2019). Less attention has been paid to settings with interventions that are continuous-

valued, so that have an associated intensity or dose. Nevertheless, estimating continuous-valued intervention responses, or simply "dose responses", is relevant in a plethora of domains of application (Farrelly et al., 2005; Calabrese, 2016; Turk, 2019; Vanderschueren et al., 2023). Compared to estimating treatment effects, dose response estimation comes with unique challenges (Schwab et al., 2019), which make this an interesting and relevant field of research.

For estimating treatment effects, controlled experiments, like randomized controlled trials (RCTs; Deaton & Cartwright, 2018), are often considered the gold standard. Yet, due to the practically infinite amount of possible doses, their application is signficantly complicated for dose response estimation (Holland-Letz & Kopp-Schneider, 2014). Even when a suitable number of units is available, such experiments might be prohibitively costly.

Alternatively, dose responses can be estimated using observational data. Yet, as with treatment effect estimation, this comes with several challenges, most importantly: a) The impossibility of observing counterfactual outcomes (Holland, 1986) and b) the presence of a dose assignment mechanism, which may have led to confounding in the data (Varadhan & Seeger, 2013). These challenges may prevent the adoption of traditional supervised learning methods (Schwab et al., 2019).

ML dose response estimators. Phenomena like observed confounding in data are well described and widely researched for treatment effect estimation (Curth et al., 2021). However, many methods presented for estimating treatment effects do not translate to the dose response setting. Methods such as matching (Ho et al., 2007), the propensity score (Rosenbaum & Rubin, 1983), representation balancing (Johansson et al., 2016), representation learning (Louizos et al., 2017) and tailored tree-based estimators (Hill, 2011; Wager & Athey, 2018) rely on the presence of two distinct groups, such as treated units and control units.

Hence, several tailored ML estimators for dose responses were proposed. Hirano & Imbens (2004) propose the Hirano-Imbens estimator (HIE), built on the idea of the generalized propensity score (GPS), extending the propensity score to continuous-valued interventions. Schwab et al. (2019) propose DRNet, a neural architecture to learn conditional-average responses to multiple intervention options with an associated dose. Their method can be combined with several balancing techniques to tackle the confounding of intervention assignment, yet not the confounding of assigned doses. Alternatively, Bica et al. (2020) propose SCIGAN, which leverages generative adversarial networks (GANs; Goodfellow et al., 2020), a method that learns how to generate counterfactual responses. Generating counterfactual responses can remove confounding in the data, enabling learning the dose response via supervised learning methods. Nie et al. (2021) propose VCNet, a generalization of DRNet that accounts for the continuity of dose responses, by training a varying coefficient network (Hastie & Tibshirani, 1993). Nie et al. (2021) study the estimation of average dose responses, by combining estimates of the conditional-average response through functional targeted regularization, a method to build doubly-robust average dose response estimators inspired by van der Laan & Rubin (2006). Wang et al. (2022) also tackle estimating the average dose response. Their estimator, ADMIT, minimizes the maximal distributional difference between units of discrete dose intervals, assuming a smooth shift in the probability distribution of covariates for units of different doses. Zhang et al. (2022) propose TransTEE, a method using the transformer architecture (Vaswani et al., 2017) to build intervention effect estimators. They claim that the attention mechanisms in these are superior in modeling observed confounders. Bellot et al. (2022) propose an architecture similar to DRNet, which uses multiple prediction heads for different interventions with a dose, and use representation balancing to overcome intervention confounding for conditional-average response estimation. Dose confounding is not addressed in their work.

Note that, even though methods such as VCNet and ADMIT were proposed for average dose response estimation, their methodologies can be used to estimate conditional-average responses. Specifically, any preceding method for average response estimation first trains a conditional-average response estimator. A simple estimator of the average response can be obtained by the law of total expectation and averaging the conditional-average estimates over a population of interest (Abrevaya et al., 2015). For a detailed description of our problem setup, see Section 3.

Benchmarking practices in ML for dose response estimation. As counterfactual outcomes cannot be observed, ML estimators for intervention responses cannot easily be evaluated using observational data

(Schwab et al., 2019; Bica et al., 2020). This is different from, for example, supervised learning (Hastie & Tibshirani, 1993). Alternatively, many researchers derive theoretical guarantees for their estimators' performance based on assumptions or use semi- or fully synthetic datasets for an empirical evaluation. Neither of these two approaches ensures good real-life performance. Datasets are usually constructed according to a synthetic data-generating process (DGP), which defines causal relationships between covariates, interventions, and outcomes. Prominent examples of such datasets for dose response estimation include the ones presented by Bica et al. (2020) and Nie et al. (2021).

Curth et al. (2021) show that most research lacks analysis of the aspects of a synthetic dataset, and that challenges in a dataset are often incompletely understood. In contrast to the preceding works, we argue that the creation of dose response estimators should be use-case driven and that we need a better understanding of the effects of different DGPs on model performance. We hence base the methodology of this work on abstractions of real-life use cases of CADR estimation, motivating the creation of our new method.

3 Problem formulation

We leverage the Neyman-Rubin potential outcomes framework (Splawa-Neyman et al., 1990; Rubin, 1974) and assume that we observe units defined by a covariate vector $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^m$ in some m-dimensional feature space. Every unit is assigned a positive dose $S \in \mathcal{S} \subset \mathbb{R}^+$. For the remainder of our paper and without loss of generality, we set $\mathcal{S} = [0, 1]$ to be the unit interval. Every unit has potential responses, or "outcomes" $Y(s) \in \mathcal{Y} \subset \mathbb{R}$, which are the responses had it received the dose s. Following Schwab et al. (2019), we call $Y(s)$ the "dose response". The expected response for a dose given a unit's covariates, the "conditional-average dose response" (CADR), is subsequently given by

$$\mu(s, \mathbf{x}) = \mathbb{E}[Y(s)|\mathbf{X} = \mathbf{x}]. \tag{1}$$

Our goal is to estimate the CADR from observational data. We expect to have access to data in the form of $\mathcal{D}_n = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^n$, where n is the total number of observed units. $\mathbf{x}_i$, s_i, and y_i are the covariates, the assigned dose, and the observed response of the i-th unit. We also refer to y_i as the "factual response". We operate under the standard fundamental problem of causal inference (Holland, 1986), so responses of a unit i to doses different from s_i, so-called "counterfactual responses", are never observed.

One challenge in CADR estimation from observational data lies in the relationships between variables $\mathbf{X}$, S, and Y. In observational data, unlike in controlled environments, doses were likely assigned according to some (partially) unknown assignment mechanism based on units' covariates. In the case of heterogenuous dose responses, such a mechanism might introduce confounding, in which the covariate vector influences both dose assignment, and the respective dose response. We illustrate the relationships between the different variables in the data in a single-world intervention graph (SWIG; Richardson & Robins, 2013) in Figure 2. To learn an unbiased estimate of the CADR, a method must hence adjust for any potential confounding (Shalit et al., 2016).

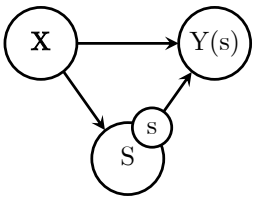


Figure 2: **Single world intervention graph** (SWIG) illustrating causal dependencies between variables in the training data. Our goal is to estimate the response of a unit to a dose s.

Finding unbiased estimates of intervention responses relies on a set of untestable assumptions (Stone, 1993). In line with preceding works, we require the following standard assumptions to hold:

Assumption 1. *(Consistency) The observed outcome Y_i for a unit i that was assigned dose s_i is the potential outcome $Y_i(s_i)$.*

Assumption 2. *(No hidden confounders) The assigned dose S is conditionally independent of the potential outcome $Y(s)$ given the covariates $\mathbf{X}$, so $\{Y(s)|s \in \mathcal{S}\} \perp\!\!\!\perp S|\mathbf{X}$*

Assumption 3. *(Overlap) Every unit has a greater-than-zero probability of receiving any dose, so $\forall s \in \mathcal{S}$: $\forall \mathbf{x} \in \mathcal{X}$ with $\mathbb{P}(\mathbf{x}) > 0 : 0 < \mathbb{P}(s|\mathbf{x}) < 1$*

Note. As indicated in Section 2, estimating the CADR is also relevant for estimating average dose responses (ADR, $\hat{\mu}(s)$). Methods for ADR estimation are typically first training a CADR estimator $\hat{\mu}(d, \mathbf{x})$ and are subsequently deriving an estimate of the ADR through the law of total expectation by calculating $\hat{\mu}(s) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(s, \mathbf{x}_i)$.

4 CBRNet: A method to tackle confounding by cluster in CADR estimation

Motivation. It is widely understood that use cases and applications should drive method selection (Luo, 2016), and that algorithm architecture plays a major role in estimator performance (Alaa & van der Schaar, 2018). This is especially true for intervention response estimators, whose performance cannot be studied on real data. We hence argue that algorithm development should be aligned more with the study of real-world applications. By now, the ML community has overly relied on a small set of manually designed benchmarking datasets (Curth et al., 2021), such as the IHDP dataset presented by Hill (2011) for treatment effect estimation, or the datasets by Bica et al. (2020) and Nie et al. (2021) for dose response estimation. While posing distinct challenges to estimators and relying on covariates acquired from real-world studies, these datasets are created using synthetic DGPs specifying intervention assignment and potential outcomes. These mechanisms are often created without further motivation, raising questions about their realisticness.

For this manuscript, we take an alternative approach and will focus on CADR estimation from clustered data with cluster-based dose assignment. Whereas in preceding works intervention and dose assignment are typically specific to a unit's covariates, we assume that similar units form clusters and that dose assignment is conditional on the cluster (cf. Figure 1). While not previously studied in the ML literature on dose response estimation, such mechanisms occur in several domains of application. Assuming that dose responses are heterogeneous, such cluster-based assignment might introduce confounding in the data, often referred to as "confounding by cluster" (CBC; Localio et al., 2002; Seaman et al., 2014; Zetterqvist et al., 2016; Berlin et al., 1999). We identified several real-life scenarios in which such mechanisms apply:

- in the pricing of loan and insurance products, companies might want to identify the effect of price on a customer's willingness to buy or renew a product, yet in accordance with established business processes similar customers were assigned to different price tiers (Yeo et al., 2002; Phillips, 2013);
- in marketing, a company might want to predict the effect of exposure to a marketing campaign, yet customers were grouped by loyalty programs and have hence received distinct amounts of discounts (Jonker et al., 2004; Ho et al., 2012);
- in medicine, practitioners might want to understand the effect of varying amounts of a medication on health outcomes, yet patients were grouped by biomarkers to receive certain doses (Schacht et al., 2014; Verhaart et al., 2014);
- in education, one might want to predict the effectiveness of certain learning programs on students' educational success, yet students were divided based on ability, leaving them exposed to similar interventions (Betts & Shkolnik, 2000);
- in public policy, policymakers might want to estimate the effects of subsidies on research organizations, yet organizations often form clusters with similar characteristics, which apply for grants jointly (Broekel et al., 2015).

In all of these scenarios, the data will likely be confounded. To learn an unbiased estimate of the CADR, we must find an estimator that is robust to such confounding. As doses are driven by the cluster, traditional supervised estimators, such as feed-forward neural networks might overfit the assignment mechanism and learn biased estimates of dose responses (Schwab et al., 2019). We propose to build a model that learns cluster-agnostic, and hence dose-agnostic representations of a unit's covariates, for unbiased CADR estimation.

Architecture. We visualize the architecture of CBRNet in Figure 3. CBRNet consists of three parts Φ, Δ, and I, based on our motivation in the paragraph above. $\Phi : \mathcal{X} \to \mathcal{R}$ is a representation learning function mapping the covariates into the representation space $\mathcal{R} \subset \mathbb{R}^n$ to learn a cluster-agnostic representation.

We use a standard feed-forward neural network for Φ, where both the number of layers and the number of hidden nodes per layer are hyperparameters. $\Delta : \mathcal{X} \times \mathcal{S} \rightarrow \{1, \ldots, k\}$ is a clustering function mapping a unit to one of k clusters by taking as input the covariates and doses of a unit. As we assume that dose assignment is conditional on the cluster, adding the dose as input to the clustering function is expected to improve performance in separating clusters with heterogeneous doses. For an overview of possible clustering algorithms, see Madhulatha (2012). We propose to use k-means (Lloyd, 1982) as a clustering function, minimizing Euclidian distances between units of a certain cluster. It is widely adopted in business and beyond (Wu, 2012) and has previously been used in treatment effect estimation (Berrevoets et al., 2020). We train Δ on all available training data and do not alter it during the training of the remaining network components (see paragraph below). $I : \mathcal{R} \times S \rightarrow \mathbb{R}$ is an inference function taking as input the covariates in representation space $\mathcal{R}$ and a dose, to learn the CADR. As for Φ, we propose I to be a feed-forward neural network with flexible hyperparameters, as similarly adopted by Shalit et al. (2016), Schwab et al. (2019), and Bica et al. (2020).

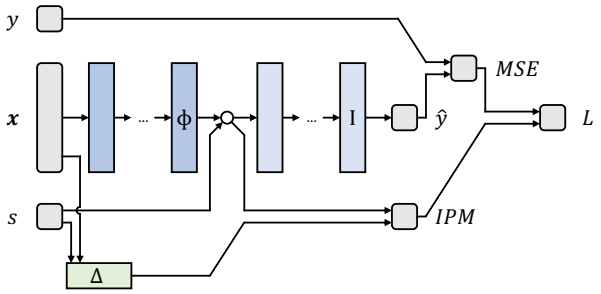


Figure 3: **Architecture of CBRNet.** The network consists of three parts. A representation learner Φ, an inference network I, and a clustering function Δ. To overcome confounding by cluster, Φ is trained to learn a dose-agnostic representation by minimizing a tailored integral probability metric (IPM) over the response space given the clusters identified by Δ. Subsequently, the inference network I is trained to learn the response of a unit to a dose by minimizing a standard mean squared error loss (MSE). For a full description of our method see Section 4.

Model selection and training. CBRNet is trained using gradient descent over the training dataset $\mathcal{D}_n = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^n$ minimizing loss L defined as

$$L(\mathbf{x}, s, y) = L_I(y, \hat{y}) + \lambda * L_\Phi(\Phi, \Delta, \mathcal{D}). \tag{2}$$

L_I is a standard mean squared error (MSE) loss

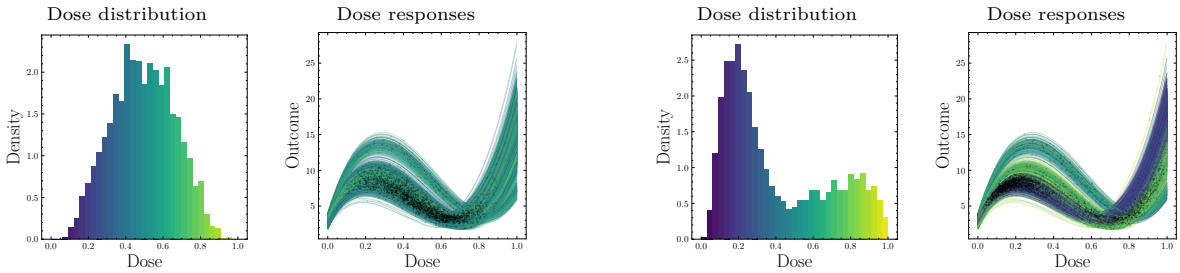
$$L_I = MSE(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3}$$

where y_i is the true outcome of unit i and $\hat{y}_i$ is the estimated outcome as calculated by the inference function I. L_Φ is a loss for the learned representation of the representation-learning network that ensures that Φ learns a cluster-agnostic representation. We rely on using integral probability metrics (IPMs) to measure the distances between the distributions of different clusters in the representation space $\mathcal{R}$. As an IMP, such as the maximum mean discrepancy (MMD; Gretton et al., 2006) or the Wasserstein distance (Kantorovich, 1960; Vaserstein, 1969), is defined for strictly two distributions, we take inspiration from Schwab et al. (2018) and calculate L_Φ as

$$L_\Phi(\Phi, \Delta, \mathcal{D}) = \frac{1}{k-1} \sum_{i=1}^{k-1} IPM\left(\{\Phi(\mathbf{x}_j)\}_{j:\Delta(\mathbf{x}_j, s_j)=i}, \{\Phi(\mathbf{x}_j)\}_{j:\Delta(\mathbf{x}_j, s_j)=k}\right). \tag{4}$$

Intuitively, we choose one base cluster k and take as regularization loss the average over the pair-wise distances between the remaining clusters and the base cluster. We set λ as a hyperparameter to balance the two losses, L_I and L_Φ. Our proposition of the regularization loss allows for a flexible choice of IPM and cluster number k, such as linear or kernel MMDs or the Wasserstein distance. We will test the impact of different choices in our empirical evaluation in the following sections.

As discussed in previous works on ML for CADR estimators, model selection and hyperparameter tuning are inherently difficult due to the unavailability of counterfactual outcomes (Schwab et al., 2019; Bica et al., 2020). For an overview, Curth & van der Schaar (2023) discuss several approaches to model selection.



(a) No confounding ($\alpha = 3$, $\beta = 0$)

(b) High level of confounding ($\alpha = 3$, $\beta = \frac{2}{3}$)

Figure 4: **Visualization of level on confounding in dry bean dataset.** Per subfigure, the left plot visualizes the dose distribution across the population and provides a color legend. The right plot visualizes the dose response space. Per dose response, the line color corresponds to the assigned dose, which is marked with a dot $(\cdot)$. In the unconfounded case (a), doses are homogeneously distributed across units. In the confounded case (b), doses are assigned conditional on clusters, as evident by the color separation of different dose responses.

Choosing one procedure over another can significantly impact response estimation and model bias. For our empirical evaluation, we propose to use a standard mean squared error over observed units in a validation set to optimize hyperparameters of CBRNet. For further detail, we refer to Section 5 and our technical implementation in Appendix D. We propose to use a validation set to optimize hyperparameters and to choose the best model in terms of the validation set MSE.

5 Experimental Evaluation

We evaluate CBRNet empirically, by comparing it to several benchmarking methods on a novel semi-synthetic dataset, the "Dry bean-DR data". The following paragraphs will discuss the creation of this data, the benchmarking methods, and the metrics used for evaluation. The dataset is available publically and has been provided with the code for this manuscript.

Data generation. To simulate clustered data and cluster-based intervention assignments, we use a semi-synthetic setup. While previous works primarily motivate the realisticness of their data by the use of covariates sampled from real-life experiments, such as the TCGA data (Cancer Genome Atlas Research Network et al., 2013), a dataset on gene sequences of cancer cells, we put a strong emphasis on the intervention assignment mechanisms as a prime factor driving confounding. Our data generation starts by taking the covariates of the dry bean dataset (Koklu & Ozkan, 2020). The dataset contains 13,611 samples of seven different types of beans, listing 16 unique features per sample derived from a computer vision analysis. We relate this data to intervention response estimation by considering a situation in which we want to understand the effect of different irrigation levels on crop yields (Goldstein et al., 2017; Dehghanisanij et al., 2022). Understanding the conditional-average responses to irrigation could lead to improved yields, yet we have to estimate these responses from observational data, stemming from established farming practices. Hence, the data likely is confounded. The data generation now proceeds in three steps:

Step 1 (Clustering): We take as clusters the types of beans in the original data and randomly aggregate them to form three clusters. For a unit i, we denote the corresponding cluster as $c_i \in \{1, 2, 3\}$. During model training and application, we expect the cluster affiliation to be unknown.

Step 2 (Dose assignment): Different from preceding datasets, we assign doses based on clusters, instead of covariates, which will lead to a more rapid shift in covariate distributions across different dose levels in the data. We refer back to Section 4 for real-world examples of such mechanisms. Every cluster j is assigned a different "modal dose" m_j, which is randomly drawn from $\{\frac{1-\beta}{2}, \frac{1}{2}, \frac{1+\beta}{2}\}$ without replacement. The parameter $\beta \in [0, 1]$ determines the variability of doses across clusters. For $\beta = 0$ the dose assignment

is homogenous across clusters. For $\beta = 1$, dose assignment is maximally heterogeneous. We set $\beta = \frac{1}{2}$ for our main experiments, and show the impact of different levels of β in Appendix C.

To add variability of doses within clusters, per unit i, we sample the individual dose s_i from a Beta-distribution

$$S_i \sim \text{Beta}(\alpha, \frac{\alpha}{m_{c_i}} + (1 - \alpha)), \tag{5}$$

as motivated by Bica et al. (2020), ensuring that the mode of S_i is m_{c_i}. The parameter α determines the variance of the Beta-distribution. For $\alpha = 0$, the dose of a unit is sampled from a uniform distribution, leading to no confounding, as the dose is unconditional on a unit's covariates. For $\alpha \to \infty$, the dose assignment is fully deterministic, and every unit in a cluster will be assigned the model dose. We can think of α as the confounding strength. We visualize the impacts of parameters α and β in Appendix A.

Step 3 (Response calculation): We define a ground truth dose response per unit, which allows for the calculation of counterfactual outcomes. The conditional-average dose response for a covariate vector $\mathbf{x}_i$ and a dose s is defined as

$$\mu(s, \mathbf{x}_i) = 10 \left(\mathbf{w}_1^\intercal \mathbf{x} + 12s \left(s - \frac{3}{4} \frac{\mathbf{w}_2^\intercal \mathbf{x}}{\mathbf{w}_3^\intercal \mathbf{x}} \right)^2 \right), \tag{6}$$

where $\mathbf{w}_l \in \mathbb{R}^{16}$ for $l \in \{1, 2, 3\}$ is a weight vector. We randomly sample half of the weights from a uniform distribution $\mathcal{U}(0, 1)$ and set the remaining weights to zero to add noise covariates (Shi et al., 2019). The individual dose response per unit is finally calculated as $y_i = \mu(s_i, \mathbf{x}_i) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$ is a random error term. The final data for different values of α and β is visualized in Figure 4 illustrating the confounding.

Benchmarks. We compare CBRNet against several relevant baselines. First, we compare it against supervised learning algorithms, namely linear regression, a regression tree, xgboost, as a state-of-the-art implementation of gradient-boosted trees and a feedforward multi-layer perceptron (MLP). Second, we compare it against DRNet (Schwab et al., 2019) and VCNet (Nie et al., 2021), two popular and competitive dose response estimators. For a discussion of these methods, see Section 2. To understand the impacts of different IPMs on model performance, we train CBRNet using the linear MMD (MMD_{lin}), a kernel MMD (MMD_{rbf}) using the radial basis function kernel, and the Wasserstein distance.

Performance metrics. We evaluate the performance of each method using the MISE metric introduced by Schwab et al. (2019), which measures the ability of a method to estimate the CADR over all units in the test data and all dose levels. For a test dataset with n units, the MISE is defined as

$$\text{MISE} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{S}} \left(\mu(s, \mathbf{x}_i) - \hat{\mu}(s, \mathbf{x}_i) \right)^2 \mathrm{d}s. \tag{7}$$

6 Empirical Results

Performance on the dry bean dataset. For a total of 17 different combinations of α and β, we generate 10 random instances of the Dry bean-DR data. We use 70% of the data for training, 10% as a validation set for hyperparameter tuning, and 20% as a test set for calculating performance metrics.

Fixing $\beta = \frac{1}{2}$, we report the performance of all nine estimators for different levels of α in Table 1. Performance under different levels of β is stated in our extended results section in Appendix C. CBRNet outperforms all benchmark methods across different levels of confounding, indicating that our proposed model architecture is well-suited to learn CADR from clustered data.

CBRNet shows higher robustness to confounding by cluster than its benchmarks when the confounding strength α increases. Irrespective of the chosen IPM for regularization, the performance of CBRNet is stable for low to moderate levels of α, only increasing for the highest value of $\alpha = 4$. For all other models, performance is strictly decreasing for increases in α.

Similar behavior can be observed for different values of β. A notable case is given for $\beta = 0$, hence when there is no variability of doses between clusters (cf. Table 2 in Appendix C). For $\beta = 0$, the data is unconfounded,

Table 1: **MISE per method on dry bean dataset** for $\beta = \frac{1}{2}$ and varying levels of α. We compare CBRNet against several benchmarking methods over varying levels of confounding by cluster. The best results are printed in **bold**, second-best results are in *italics*. Results for different values of α can be found in Appendix C.

$\beta = \frac{1}{2}$

	α			
Model	1.0	2.0	3.0	4.0
Linear Regression	3.32 ± 0.03	3.40 ± 0.06	3.30 ± 0.02	3.32 ± 0.03
CART	1.31 ± 0.06	1.40 ± 0.09	1.56 ± 0.15	1.53 ± 0.13
xgboost	0.91 ± 0.07	1.16 ± 0.09	1.15 ± 0.07	1.28 ± 0.04
MLP	0.54 ± 0.04	0.67 ± 0.08	0.83 ± 0.09	1.06 ± 0.08
DRNet	0.62 ± 0.03	0.78 ± 0.13	0.82 ± 0.07	0.98 ± 0.06
VCNet	0.75 ± 0.11	1.07 ± 0.10	1.18 ± 0.11	1.35 ± 0.07
CBRNet(MMD_{lin})	*0.44* ± 0.09	*0.47* ± 0.12	**0.44** ± 0.07	*0.63* ± 0.15
CBRNet(MMD_{rbf})	0.45 ± 0.10	0.53 ± 0.14	0.50 ± 0.10	0.71 ± 0.17
CBRNet(Wass)	**0.43** ± 0.09	**0.46** ± 0.10	*0.45* ± 0.08	**0.55** ± 0.11

as doses are sampled from the same distribution for all units and all clusters. So, CBRNet also outperforms the benchmark methods when learning CADR from clustered, but unconfounded data.

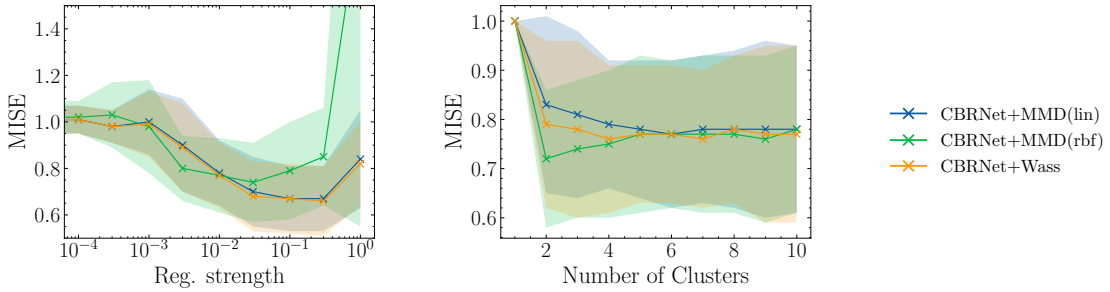
Hyperparameter sensitivity We further investigate the sources of gain of our proposed architecture by testing the impact of two critical hyperparameters: (a) The strength of the regularization λ, and (b) the number of clusters k identified by the k-means clustering. The results of those experiments are summarized in Figure 5. For both experiments, we fix the hyperparameters of CBRNet and iterate over varying levels of λ and k. To stabilize results, for every parameter configuration, we generate 10 random instances of the Dry bean-DR data with $\alpha = \frac{2}{3}$ and $\beta = 3$.

In Appendix B we visualize the impact of the IMP regularization on the representation space. We find a strictly positive impact on model performance for low to moderate levels of regularization ($\lambda < 0.1$). These results support the positive impact of regularizing for differences in cluster distributions in the representation space. These results are also independent of the chosen IPM, with MMDs and Wasserstein distance performing comparably. Only for an increased level of regularization, does the predictive performance of CBRNet decrease. This indicates that an overregularization is possible

The performance of CBRNet appears to be robust to the specified number of clusters k, with performance converging on the Dry bean-DR data for larger k. We hence expect our method to have favorable asymptotical properties for unclustered data, as long as the dose assignment can be approximated sufficiently well by a large enough number of clusters and if sufficient data is available to approximate the IPMs. In applications, the number of clusters can be selected together with domain experts, or by visual analysis.

Performance on established datasets To judge the performance of CBRNet on unclustered data, we run experiments on previously established benchmarking datasets proposed by Bica et al. (2020)[2] and Nie et al. (2021). CBRNet performs competitively, beating benchmarks on three of the four datasets. However, further decomposition is needed to understand the precise challenges in these datasets to attribute performance to either the general architecture of CBRNet, or the IPM regularization. We provide the full results on these datasets in Table 5 in Appendix C.

[2]The DGP proposed by Bica et al. (2020) can incorporate up to three different interventions, each with an associated dose. To comply with the architecture of CBRNet, we generate responses to only a single intervention.



(a) Impact of reg. strength

(b) Impact of num. clusters

Figure 5: **Hyperparameter robustness.** (a) Moderate levels of regularization improve model performance. Regularizing by the MMD with a radial basis function kernel (rbf) appears to overregularize earlier than for linear MMD and Wasserstein distance. (b) We see robustness to the number of clusters. Performance converges for all IPMs.

7 Conclusion

We studied the estimation of conditional-average dose responses (CADR) from clustered observational data. Practitioners face clustered data in several real-life scenarios, covering applications in, e.g., business, economics, and healthcare (cf. Section 4). In such scenarios, dose assignment is often cluster-specific, potentially leading to confounding. This confounding mechanism is distinct from scenarios previously studied in the ML literature on intervention response estimation and has impacts on the validity of assumptions that are often made beyond standard ones necessary for applying causal inference (cf. Section 2).

To enable thorough testing of estimators under the presence of such data, we created a novel dataset, the "Dry bean-DR data", based on the covariates initially presented by Koklu & Ozkan (2020). Additionally, we proposed a new ML estimator, CBRNet, which leverages representation balancing to learn cluster-agnostic representations of training data to prevent confounding biases. Our results reveal that traditional supervised estimators and ML estimators tailored to CADR estimation suffer from confounding by cluster. In comparison, CBRNet improved upon the performance of these methods, indicating our approach's appropriateness. Further experiments using established datasets revealed that CBRNet can similarly achieve competitive performance on unclustered data.

Our research has shown that understanding the context and unique characteristics of a CADR estimation problem is imperative. Variations in the underlying data-generating process of observational data can have adverse effects on model performance. To ensure good performance, especially given the inherent difficulty of model selection in causal inference, practitioners must check the validity of any method-specific assumptions, for example, by involving domain experts (Pearl, 2022).

Our implementation of CBRNet is available online for practitioners and fellow researchers to build upon (cf. Appendix D). The architecture of our method was motivated by the analysis of real-world scenarios, leveraging and extending the concept of representation balancing, first proposed in the literature on ML for causal inference by Shalit et al. (2016). We highlight three potential future research directions: (1) A theoretical analysis of CBRNet and asymptotical guarantees of model performance in clustered and unclustered data. Especially, an analysis of the necessary amount of data for effective regularization could benefit the adoption of our model, as the IPMs for regularization are approximated empirically. (2) Further research on model selection and hyperparameter tuning. While using a validation set MSE can be a practical first guess (Curth & van der Schaar, 2023), alternative approaches could further add performance guarantees when training models on real-life data. This research is not specific to CBRNet, but is relevant to the wider related literature. (3) Finally, our research is specific to interventions with an associated dose, hence extending our experiments to binary-valued interventions could be a promising and insightful direction, especially given the limited number of benchmarking datasets for treatment effect estimation (Curth et al., 2021).

References

Jason Abrevaya, Yu-Chin Hsu, and Robert P. Lieli. Estimating conditional average treatment effects. *Journal of Business and Economic Statistics*, 33(4):485–505, October 2015. ISSN 1537-2707. doi: 10.1080/07350015.2014.975555.

Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138. PMLR, 2018.

Alexis Bellot, Anish Dhir, and Giulia Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage, May 2022.

Jesse A. Berlin, Stephen E. Kimmel, Thomas R. Ten Have, and Mary D. Sammel. An Empirical Comparison of Several Clustered Data Approaches Under Confounding Due to Cluster Effects in the Analysis of Complications of Coronary Angioplasty. *Biometrics*, 55(2):470–476, 1999. ISSN 1541-0420. doi: 10.1111/j.0006-341X.1999.00470.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.1999.00470.x.

Jeroen Berrevoets, James Jordon, Ioana Bica, and Mihaela van der Schaar. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. *NIPS*, 33:20037–20050, 2020.

Julian R. Betts and Jamie L. Shkolnik. The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, 19(1):1–15, February 2000. ISSN 0272-7757. doi: 10.1016/s0272-7757(98)00044-2.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems (2020)*, February 2020. doi: 10.48550/ARXIV.2002.12326.

Tom Broekel, Dirk Fornahl, and Andrea Morrison. Another cluster premium: Innovation subsidies and R&D collaboration networks. *Research Policy*, 44(8):1431–1444, October 2015. ISSN 0048-7333. doi: 10.1016/j.respol.2015.05.002.

Edward Calabrese. The emergence of the dose–response concept in biology and medicine. *International Journal of Molecular Sciences*, 17(12):2034, December 2016. ISSN 1422-0067. doi: 10.3390/ijms17122034.

Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013. ISSN 1546-1718. doi: 10.1038/ng.2764.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system, August 2016.

Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation, February 2023.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=FQLzQqGEAH.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med*, 210:2–21, 2018.

Hossein Dehghanisanij, Hojjat Emami, Somayeh Emami, and Vahid Rezaverdinejad. A hybrid machine learning approach for estimating the water-use efficiency and yield in agriculture. *Scientific Reports*, 12 (1), April 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-10844-2.

William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, Cinjon, and Anton Bakhtin. Pytorch lightning, 2020. URL https://www.pytorchlightning.ai.

Matthew C. Farrelly, Kevin C. Davis, M. Lyndon Haviland, Peter Messeri, and Cheryl G. Healton. Evidence of a dose—response relationship between "truth" antismoking ads and youth smoking prevalence. *American Journal of Public Health*, 95(3):425–431, March 2005. ISSN 1541-0048. doi: 10.2105/ajph.2004.049692.

Emil Frei and George P. Canellos. Dose: A critical factor in cancer chemotherapy. *The American Journal of Medicine*, 69(4):585–594, October 1980. ISSN 0002-9343. doi: 10.1016/0002-9343(80)90472-6. URL https://www.sciencedirect.com/science/article/pii/0002934380904726.

Anat Goldstein, Lior Fink, Amit Meitin, Shiran Bohadana, Oscar Lutenberg, and Gilad Ravid. Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. *Precision Agriculture*, 19(3):421–444, May 2017. ISSN 1573-1618. doi: 10.1007/s11119-017-9527-4.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, October 2020. ISSN 0001-0782. doi: 10.1145/3422622. URL https://dl.acm.org/doi/10.1145/3422622.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *NIPS*, 19, 2006.

Sebastien Haneuse. Distinguishing Selection Bias and Confounding Bias in Comparative Effectiveness Research. *Medical Care*, 54(4):e23–29, April 2016. ISSN 1537-1948. doi: 10.1097/MLR.0000000000000011.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, September 1993. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1993.tb01939.x.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, January 2011. ISSN 1537-2715. doi: 10.1198/jcgs.2010.08162.

Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments, July 2004. ISSN 1940-6347.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007. ISSN 1476-4989. doi: 10.1093/pan/mpl013.

G. T. S. Ho, W. H. Ip, C. K. M. Lee, and W. L. Mou. Customer grouping for better resources allocation using GA based clustering technique. *Expert Systems with Applications*, 39(2):1979–1987, February 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2011.08.045.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81 (396):945–960, December 1986. ISSN 0162-1459. doi: 10.1080/01621459.1986.10478354. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

Tim Holland-Letz and Annette Kopp-Schneider. Optimal experimental designs for dose–response studies with continuous endpoints. *Archives of Toxicology*, 89(11):2059–2068, August 2014. ISSN 1432-0738. doi: 10.1007/s00204-014-1335-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4655015/pdf/204_2014_Article_1335.pdf.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference, May 2016.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects, January 2020.

Jedid-Jah Jonker, Nanda Piersma, and Dirk Van den Poel. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications*, 27(2):159–168, August 2004. ISSN 0957-4174. doi: 10.1016/j.eswa.2004.01.010. URL https://www.sciencedirect.com/science/article/pii/S095741740400003X.

L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6 (4):366–422, July 1960. ISSN 1526-5501. doi: 10.1287/mnsc.6.4.366.

Murat Koklu and Ilker Ali Ozkan. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174:105507, July 2020. ISSN 0168-1699. doi: 10.1016/j.compag.2020.105507. URL https://www.sciencedirect.com/science/article/pii/S0168169919311573.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Trans Inf Theory*, 28(2):129–137, 1982.

A. Russell Localio, Jesse A. Berlin, and Thomas R. Ten Have. Confounding Due to Cluster in Multicenter Studies—Causes and Cures. *Health Services and Outcomes Research Methodology*, 3(3):195–210, December 2002. ISSN 1572-9400. doi: 10.1023/A:1025897811143. URL https://doi.org/10.1023/A:1025897811143.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models, May 2017.

Gang Luo. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):18, May 2016. ISSN 2192-6670. doi: 10.1007/s13721-016-0125-6. URL https://doi.org/10.1007/s13721-016-0125-6.

T. Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments, March 2021.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects, December 2017.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, et al. Automatic differentiation in PyTorch. In *NIPS-W*, pp. 1–4, 2017.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, reprinted with corrections edition, 2022. ISBN 9780521895606. Hier auch später erschienene, unveränderte Nachdrucke.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Robert Phillips. Optimizing prices for consumer credit. *Journal of Revenue and Pricing Management*, 12: 360–377, 2013.

Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, oct 1974. doi: 10.1037/h0037350.

Alexander Schacht, Philip Gorwood, Philip Boyce, Ayal Schaffer, and Hernan Picard. Depression symptom clusters and their predictive value for treatment outcomes: Results from an individual patient data meta-analysis of duloxetine trials. *Journal of Psychiatric Research*, 53:54–61, June 2014. ISSN 0022-3956. doi: 10.1016/j.jpsychires.2014.02.001. URL https://www.sciencedirect.com/science/article/pii/S0022395614000314.

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks, October 2018.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5612–5619, apr 2019. doi: 10.1609/aaai.v34i04.6014.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Shaun Seaman, Menelaos Pavlou, and Andrew Copas. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine*, 33(30):5371–5387, 2014. ISSN 1097-0258. doi: 10.1002/sim.6277. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6277.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, June 2016.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects, June 2019.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237. URL http://www.jstor.org/stable/2245382.

Richard Stone. The Assumptions on Which Causal Inferences Rest. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):455–466, January 1993. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1993.tb01915.x. URL https://doi.org/10.1111/j.2517-6161.1993.tb01915.x.

Jonathan M. Turk. Estimating the impact of developmental education on associate degree completion: A dose–response approach. *Research in Higher Education*, 60(8):1090–1112, March 2019. ISSN 1573-188X. doi: 10.1007/s11162-019-09549-9.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), January 2006. ISSN 1557-4679. doi: 10.2202/1557-4679.1043.

Guido Van Rossum, Fred L. Drake, et al. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Toon Vanderschueren, Robert Boute, Tim Verdonck, Bart Baesens, and Wouter Verbeke. Optimizing the preventive maintenance frequency with causal machine learning. *International Journal of Production Economics*, 258:108798, apr 2023. doi: 10.1016/j.ijpe.2023.108798.

Ravi Varadhan and John D. Seeger. Estimation and Reporting of Heterogeneity of Treatment Effects. In *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Agency for Healthcare Research and Quality (US), January 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK126190/pdf/Bookshelf_NBK126190.pdf.

Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredači Informacii*, 5(3):64–72, 1969.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, June 2017.

Brian G. Vegetabile. On the distinction between "conditional average treatment effects" (cate) and "individual treatment effects" (ite) under ignorability assumptions, August 2021.

René F. Verhaart, Valerio Fortunati, Gerda M. Verduijn, Theo van Walsum, Jifke F. Veenland, and Margarethus M. Paulides. CT-based patient modeling for head and neck hyperthermia treatment planning: Manual versus automatic normal-tissue-segmentation. *Radiotherapy and Oncology*, 111(1):158–163, April 2014. ISSN 0167-8140. doi: 10.1016/j.radonc.2014.01.027. URL https://www.sciencedirect.com/science/article/pii/S0167814014001005.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, June 2018. ISSN 1537-274X. doi: 10.1080/01621459.2017.1319839.

Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. Generalization bounds for estimating causal effects of continuous treatments. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8605–8617. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/390bb66a088d37f62ee9fb779c5953c2-Paper-Conference.pdf.

Junjie Wu. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.

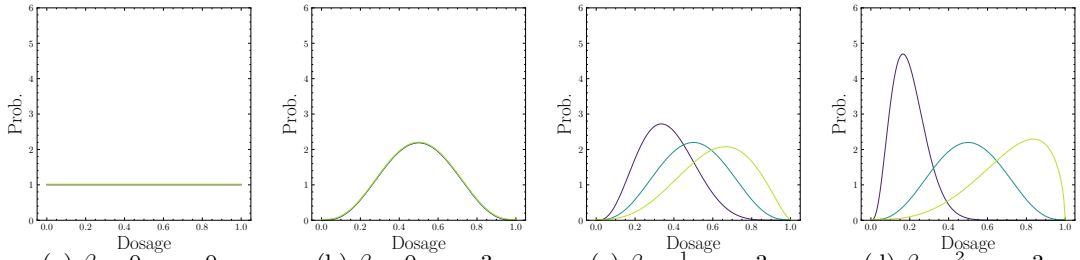
A C Yeo, K A Smith, R J Willis, and M. Brooks. A mathematical programming approach to optimise insurance premium pricing within a data mining framework. *Journal of the Operational Research Society*, 53(11):1197–1203, November 2002. ISSN 1476-9360. doi: 10.1057/palgrave.jors.2601413. URL https://doi.org/10.1057/palgrave.jors.2601413.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Johan Zetterqvist, Stijn Vansteelandt, Yudi Pawitan, and Arvid Sjölander. Doubly robust methods for handling confounding by cluster. *Biostatistics*, 17(2):264–276, April 2016. ISSN 1465-4644. doi: 10.1093/biostatistics/kxv041. URL https://doi.org/10.1093/biostatistics/kxv041.

Yi-Fan Zhang, Hanlin Zhang, Zachary C. Lipton, Li Erran Li, and Eric P. Xing. Exploring transformer backbones for heterogeneous treatment effect estimation, February 2022.

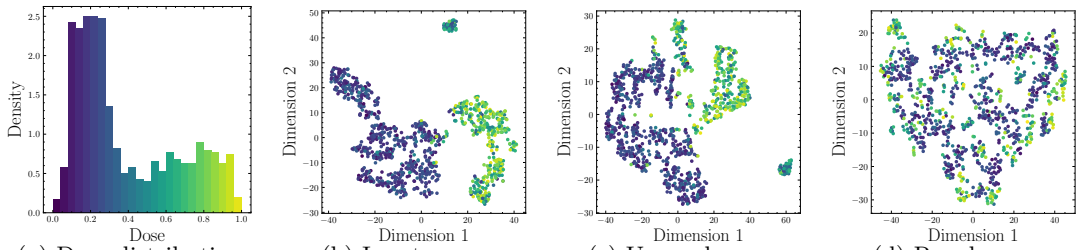
A Visualization of confounding mechanism



(a) $\beta = 0$, $\alpha = 0$ (b) $\beta = 0$, $\alpha = 3$ (c) $\beta = \frac{1}{3}$, $\alpha = 3$ (d) $\beta = \frac{2}{3}$, $\alpha = 3$

Figure 6: **Visualization of dose distributions per cluster.** Subfigures visualize the dose distributions of different clusters for varying levels of confounding. Parameter β determines the difference in doses between individual clusters. Parameter α determines the variability of doses within a cluster.

B Effect of regularization on covariate representation



(a) Dose distribution (b) Input space (c) Unregul. rep. (d) Regul. rep.

Figure 7: **Regularizing impact of integral probability metric (IPM).** We visualize the regularizing impact of the IPM. Figure (a) visualizes the density of different doses in the test data and provides a color legend. The preceding figures are two-dimensional t-SNE plots of (b) the input space in $\mathcal{X}$, (c) the learned representation of CBRNet without regularization, and (d) the learned representation with regularization by the MMD with a radial basis function kernel. The regularization effectively removes the clustering by dose level.

C Extended results

Table 2: **MISE per method on dry bean dataset** for $\beta = 0$ and varying levels of α.

$\beta = 0$

	α				
Model	0.0	1.0	2.0	3.0	4.0
Linear Regression	2.83 ± 0.01	3.27 ± 0.03	3.80 ± 0.03	4.19 ± 0.04	4.44 ± 0.04
CART	0.86 ± 0.03	1.04 ± 0.03	1.44 ± 0.07	1.93 ± 0.10	2.18 ± 0.11
xgboost	0.55 ± 0.03	0.69 ± 0.02	1.11 ± 0.08	1.58 ± 0.07	1.88 ± 0.10
MLP	2.83 ± 0.01	0.54 ± 0.08	1.17 ± 0.09	1.98 ± 0.05	2.52 ± 0.09
DRNet	0.41 ± 0.03	0.60 ± 0.06	0.99 ± 0.12	1.74 ± 0.19	1.92 ± 0.23
VCNet	0.34 ± 0.04	0.69 ± 0.04	1.04 ± 0.09	1.41 ± 0.09	1.69 ± 0.17
CBRNet(MMD_{lin})	**0.25** ± 0.04	*0.35* ± 0.05	**0.41** ± 0.10	0.64 ± 0.08	*0.88* ± 0.14
CBRNet(MMD_{rbf})	*0.26* ± 0.04	**0.34** ± 0.05	*0.42* ± 0.07	**0.61** ± 0.08	0.92 ± 0.09
CBRNet(Wass)	0.28 ± 0.05	**0.34** ± 0.05	*0.42* ± 0.10	*0.62* ± 0.08	**0.86** ± 0.16

Table 3: **MISE per method on dry bean dataset** for $\beta = \frac{1}{4}$ and varying levels of α.

$\beta = \frac{1}{4}$

	α			
Model	1.0	2.0	3.0	4.0
Linear Regression	3.43 ± 0.03	3.84 ± 0.04	4.06 ± 0.03	4.35 ± 0.04
CART	1.17 ± 0.09	1.47 ± 0.11	1.74 ± 0.08	2.04 ± 0.21
xgboost	0.77 ± 0.07	1.04 ± 0.07	1.37 ± 0.09	1.66 ± 0.13
MLP	0.57 ± 0.06	1.01 ± 0.11	1.46 ± 0.13	2.14 ± 0.13
DRNet	0.72 ± 0.07	*0.96* ± 0.07	1.30 ± 0.16	1.54 ± 0.20
VCNet	0.68 ± 0.05	1.16 ± 0.08	1.48 ± 0.16	1.83 ± 0.12
CBRNet(MMD_{lin})	**0.41** ± 0.09	**0.57** ± 0.17	*0.71* ± 0.17	*0.79* ± 0.22
CBRNet(MMD_{rbf})	*0.42* ± 0.11	**0.57** ± 0.15	0.75 ± 0.26	0.87 ± 0.15
CBRNet(Wass)	0.44 ± 0.08	**0.57** ± 0.15	**0.69** ± 0.28	**0.74** ± 0.17

Table 4: **MISE per method on dry bean dataset** for $\beta = \frac{3}{4}$ and varying levels of α.

$\beta = \frac{3}{4}$

	α			
Model	1.0	2.0	3.0	4.0
Linear Regression	2.91 ± 0.01	2.92 ± 0.02	3.01 ± 0.03	3.01 ± 0.02
CART	1.24 ± 0.11	1.31 ± 0.08	1.32 ± 0.10	1.43 ± 0.11
xgboost	1.04 ± 0.05	1.04 ± 0.08	1.00 ± 0.06	1.09 ± 0.06
MLP	1.19 ± 0.09	1.19 ± 0.1	1.64 ± 0.11	1.57 ± 0.15
DRNet	0.60 ± 0.07	0.59 ± 0.06	0.71 ± 0.05	0.70 ± 0.08
VCNet	0.55 ± 0.10	0.62 ± 0.07	0.60 ± 0.05	0.66 ± 0.07
CBRNet(MMD_{lin})	*0.35* ± 0.07	*0.39* ± 0.11	**0.39** ± 0.09	**0.45** ± 0.06
CBRNet(MMD_{rbf})	0.43 ± 0.11	0.49 ± 0.18	*0.42* ± 0.06	*0.51* ± 0.13
CBRNet(Wass)	**0.34** ± 0.06	**0.38** ± 0.10	*0.42* ± 0.08	**0.45** ± 0.07

Table 5: **MISE on established benchmark datasets.** CBRNet performs competitively against benchmark methods across established datasets, achieving state-of-the-art performance on three out of four datasets, and improving over a standard MLP on all datasets.

	Dataset			
Model	TCGA-2	IHDP-1	News-2	Synth-1
MLP	2.01 ± 0.1	2.79 ± 0.09	1.22 ± 0.12	1.60 ± 1.22
DRNet	0.35 ± 0.05	2.54 ± 0.08	*0.81* ± 0.07	1.52 ± 1.06
VCNet	0.68 ± 0.30	*1.42* ± 0.19	**0.78** ± 0.06	*0.90* ± 0.55
CBRNet(MMD_{lin})	**0.24** ± 0.07	1.60 ± 0.33	1.12 ± 0.06	0.99 ± 0.66
CBRNet(MMD_{rbf})	*0.26* ± 0.09	1.56 ± 0.17	1.10 ± 0.05	0.99 ± 0.61
CBRNet(Wass)	**0.24** ± 0.08	**1.40** ± 0.25	1.10 ± 0.05	**0.89** ± 0.56

D Implementation, reproducibility, and hyperparameter optimization

Our experiments are written in Python 3.10 (Van Rossum et al., 1995) and were executed on an Apple M2 Pro SoC with 10 CPU cores, 16 GPU cores, and 16 GB of shared memory. The system needs approximately one day for the iterative execution of all experiments. We aim for maximal reproducibility. All methods considered in our manuscript are consistently implemented in an sklearn style. The code to reproduce all experiments, results, and figures paper can be found online via https://anonymous.4open.science/r/CBRNet-600C/.

For VCNet, we build on the original implementation provided by Nie et al. (2021) (https://github.com/lushleaf/varying-coefficient-net-with-functional-tr). All remaining neural network architectures were implemented in PyTorch (Paszke et al., 2017) using Lightning (Falcon et al., 2020). Xgboost is implemented using the `xgboost` library (Chen & Guestrin, 2016). CART models were implemented using the `Scikit-Learn` library (Pedregosa et al., 2011). Linear regression was implemented using the `statsmodels` library (Seabold & Perktold, 2010).

For the TCGA-2 dataset, linear regression models were trained using the first 50 principal components of the covariate matrix to reduce computational complexity.

Hyperparameter optimization. Results are not to be compared to the original papers, as the optimization scheme and parameter search ranges differ from the original records. If not specified differently, the remaining hyperparameters are set to match the specifications of the original authors.

Table 6: Hyperparameter search range for Linear Regression:

Parameter	Values
Penalty	$\{Elastic\ net, Lasso, None\}$

Table 7: Hyperparameter search range for CART:

Parameter	Values
Max depth	$\{5, 15, None\}$
Min sample split	$\{2, 5, 20\}$
Min samples per leaf	$\{1, 5, 10\}$
Max features per split	$\{None, \sqrt{p(\mathbf{x})}\}$
Splitting criterion	$\{Gini\}$

Table 8: Hyperparameter search range for xgboost:

Parameter	Values
Learning rate	$\{0.01, 0.1, 0.2\}$
Max depth	$\{3, 5, 7, 9\}$
Subsample	$\{0.5, 0.7, 1.0\}$
Min child weight	$\{1, 3, 5\}$
Gamma	$\{0.0, 0.1, 0.2\}$
Columns sampled per tree	$\{0.3, 0.5, 0.7\}$

Table 9: Hyperparameter search range for MLP:

Parameter	Values
Learning rate	$\{0.0001, 0.001\}$
L2 regularization	$\{0.0, 0.1\}$
Batch size	$\{64, 128\}$
Hidden size	$\{32, 48\}$
Num steps	$\{5000\}$
Num layers	$\{2\}$
Optimizer	$\{Adam\}$

Table 10: Hyperparameter search range for DRNet:

Parameter	Values
Learning rate	$\{0.0001, 0.001\}$
L2 regularization	$\{0.0, 0.1\}$
Batch size	$\{64, 128\}$
Hidden size	$\{32, 48\}$
Num dose strata	$\{10\}$
Num steps	$\{5000\}$
Num layers	$\{2\}$
Optimizer	$\{Adam\}$

Table 11: Hyperparameter search range for VCNet:

Parameter	Values
Learning rate	$\{0.001, 0.01\}$
Batch size	$\{128, 256\}$
Hidden size	$\{32\}$
Num steps	$\{5000\}$
Optimizer	$\{Adam\}$

Table 12: Hyperparameter search range for CBRNet:

Parameter	Values
Learning rate	$\{0.001, 0.01\}$
Batch size	$\{128, 256\}$
Hidden size	$\{32\}$
Num steps	$\{5000\}$
Optimizer	$\{Adam\}$
IPM Regularization	$\{0.001, 0.01, 0.1\}$