Conditional average treatment effect estimation with treatment offset models

Wouter A.C. van Amsterdam
Babylon Health
w.a.c.vanamsterdam@gmail.com

Rajesh Ranganath
Courant Institute of Mathematical Sciences
New York University
rajeshr@cims.nyu.edu

Abstract

In healthcare, treatment effect estimates from randomized controlled trials for binary outcomes are often reported on a relative scale using the odds-ratio. To weigh potential benefits and harms of treatment this odds-ratio has to be translated to a difference in absolute risk, preferably on an individual patient level. Under the assumption that the *relative* treatment effect is fixed, treatments have widely varying effects on an *absolute* risk scale for patients with different *untreated* risk. We investigate whether an estimate of the *relative* treatment effect from randomized trials can be exploited for estimating the treatment effect on an absolute risk scale conditional on covariates in the presence of unobserved confounding using *treatment offset models*. We first demonstrate for a simple example that this is not the case. We then investigate the magnitude of the resulting confounding bias using numerical experiments based on a binary confounder. We find that for virtually all plausible confounding magnitudes estimating the conditional average treatment effect using offset models is more accurate than assuming a single absolute treatment effect whenever the observed conditional association between the covariates and the outcome in the observational data is large enough. Finally, we evaluate a neural network-based offset model on a task with real-world medical images and simulated outcome data and find that the offset model performs well.

1 Introduction

In healthcare, treatment effect estimates from randomized controlled trials for binary outcomes are often reported on a relative scale using the odds-ratio (e.g. Furie et al. (2020); Simonovich et al. (2021); Lean et al. (2018)). This implicitly assumes that this single odds-ratio holds for all patients included in the trial population. To weigh potential benefits and harms of treatment for an individual patient, this odds-ratio has to be translated to a difference in absolute risk, preferably conditional on characteristics of the patient. Under the assumption that the *relative treatment effect* is fixed, treatments have widely varying effects on an *absolute risk scale* depending on the *untreated risk* of a patient (for an illustration, see Figure 1).

For instance, assume that a cholesterol lowering drug reduces the risk of cardiovascular death within the next 10 years by 50%. A 60 year-old male smoker with hypertension and raised cholesterol has an untreated risk of cardiovascular death of 40% and should expect a reduction in risk of 20% points. A 50 year-old female without hypertension has an untreated risk of under 1% and will have a less than 0.5% points reduction in risk. Given these widely different effects on an absolute probability scale, one may recommend the cholesterol lowering drug to the 60-year old male but not the 50-year old female. Models that predict the difference in outcomes between two treatments conditional on covariates are conditional average treatment effect (CATE) models.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

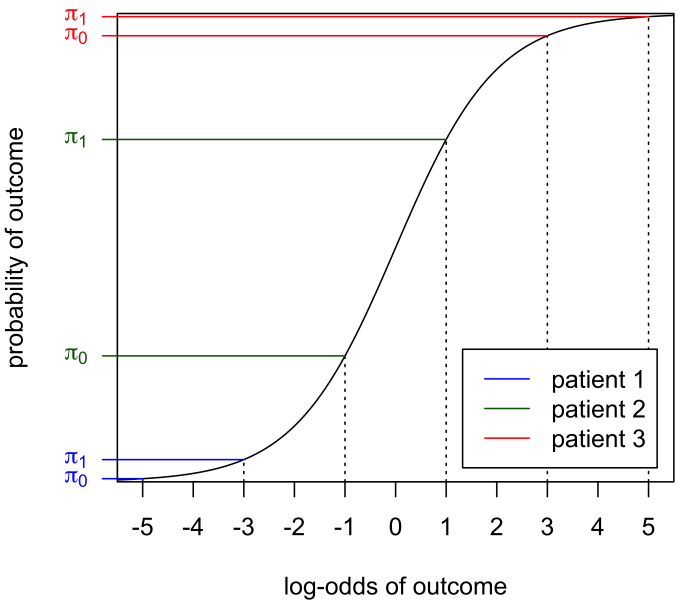

Figure 1: Treatments with a fixed effect on a log-odds scale have varying effects on an *absolute risk* scale (CATE $:= \pi_1 - \pi_0$) depending on the *untreated risk* (π_0) of the patient.

Several previous studies on breast cancer patients used the assumption of a fixed relative treatment effect to develop CATE models from observational data using *treatment offset models* (Candido dos Reis et al., 2017; Ravdin et al., 2001; Alaa et al., 2021). The purpose of these models is to help determine whether it is beneficial to give chemotherapy after surgical removal of the breast tumor. Using treatment effect estimates from randomized controlled trials on a relative scale, these models estimate the absolute survival probability under chemotherapy or no chemotherapy. After the models were found to be accurate in observational validation studies, treatment guidelines acknowledged a place for these models in clinical decision making (Cardoso et al., 2019; Gradishar, 2021). In these models it is implicitly assumed that a fixed relative treatment effect allows one to estimate a CATE model from observational data even in the presence of unobserved confounding. However, whether this is correct has not been verified rigorously.

In addition, in recent years there has been much interest in the application of neural networks to unstructured data such as medical images for improving prognosis predictions. Though there have been studies on estimating treatment effects using neural networks (e.g. Shalit et al. (2017); Shi et al. (2019)), including from medical images (van Amsterdam et al., 2019), the fixed treatment effect assumption has not been applied to neural network based models for unstructured data.

In this work, we evaluate the validity of the assumption that a fixed relative treatment effects allows for CATE estimation in the presence of unobserved confounding using offset models. We find that this is not the case, but in numerical experiments the bias was low enough that using offset models still leads to better estimation of the individual risk reduction associated with treatment compared with the baseline of assuming a single risk difference for all patients. We then use offset models on a task with medical images using convolutional neural networks and find that the models work well.

2 Methods

We consider models that predict the absolute difference in probability of a binary outcome y under two possible treatments conditional on some pre-treatment covariate vector $\boldsymbol{w}$. This is the conditional average treatment effect (CATE), conditional on $\boldsymbol{w}$:

$$\text{CATE}(\boldsymbol{w}) := p(y = 1|\text{do}(x = 1), \boldsymbol{w}) - p(y = 1|\text{do}(x = 0), \boldsymbol{w}) \tag{1}$$

We assume that the relative treatment effect β_x on a log odds-ratio scale is known from randomized trials. Odds are defined relative to a probability π as $\text{odds}(\pi) = \frac{\pi}{1-\pi}$. The odds-ratio of two probabilities π_0, π_1 is defined as $\text{OR}(\pi_0, \pi_1) := \text{odds}(\pi_0)/\text{odds}(\pi_1)$. Writing $\pi_{x'}(\boldsymbol{w'}) = p(y = 1|\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'})$, the assumption that the odds-ratio of the outcome under treatment or no treatment is constant implies that for any two values $\boldsymbol{w}_0, \boldsymbol{w}_1$ of $\boldsymbol{w}$, $\text{OR}(\pi_0(\boldsymbol{w}_0), \pi_1(\boldsymbol{w}_0)) = \text{OR}(\pi_0(\boldsymbol{w}_1), \pi_1(\boldsymbol{w}_1))$. Or equivalently, the log-odds of the outcome under $\text{do}(x = 1)$ versus $\text{do}(x = 0)$ differ by a constant β_x. Introducing $\eta(x, \boldsymbol{w})$ as the log-odds of the outcome, the assumption implies that for each $x', \boldsymbol{w'}$:

$$\eta(\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'}) = \beta_0(\boldsymbol{w'}) + \beta_x x' \tag{2}$$

Where $\beta_0(\boldsymbol{w})$ is the log-odds of the outcome in the group with $\text{do}(x = 0)$ as a function of $\boldsymbol{w}$. Denote $\sigma(x) = \frac{1}{1+e^{-x}}$ the sigmoid function such that $\sigma(\log \text{odds}(\pi)) = \pi$, we can now write the CATE in terms of η:

$$\text{CATE}(\boldsymbol{w}) = \sigma(\eta(1, \boldsymbol{w})) - \sigma(\eta(0, \boldsymbol{w})) \tag{3}$$

A fixed term in a model that is not estimated from data is called an *offset* term (Watson, 2007). We therefore refer to models of the from of Equation 2 as *treatment offset models* or *offset models* for short.

2.1 Estimating offset models

Given the form of the log-odds ratio in Equation 2 and the assumption that β_x is given a priori, offset models can be estimated with likelihood based approaches by specifying a parametric model for $\hat{\beta}_0(\boldsymbol{w}, \boldsymbol{\theta})$. The full model is then given by $\hat{\eta}(x, \boldsymbol{w}, \boldsymbol{\theta}) = \hat{\beta}_0(\boldsymbol{w}, \boldsymbol{\theta}) + \beta_x x$. In the case of logistic regression, $\hat{\beta}_0(\boldsymbol{w}, \boldsymbol{\theta}) = \theta_0 + \boldsymbol{\theta}_{\boldsymbol{w}}\boldsymbol{w}$. However, $\hat{\beta}_0(\boldsymbol{w}, \boldsymbol{\theta})$ may also be a more flexible function, for example a convolutional neural network when $\boldsymbol{w}$ is an image.

2.2 Identification

We assume the Acyclic Mixed Directed Graph (AMDG) with observed covariate W and unobserved confounder U presented in Figure 2.

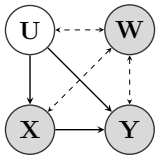

Figure 2: Acyclic Mixed Directed Graph with observed nodes X, W, Y and unobserved confounder U

To prove that the CATE is identified it is sufficient to prove that $p(y = 1|\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'})$ is identified for all $x', \boldsymbol{w'}$. Due to the unobserved confounder U in the AMDG, $p(y = 1|\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'})$ is not identifiable from observational data *in general*. The question is whether the additional assumption of a fixed relative treatment effect as stated in Equation 2 is sufficient for $\boldsymbol{w}$-conditional causal effect identification from observational data when U is not observed. The assumption implies that our query is identified when $\beta_0(\boldsymbol{w})$ is identified, as $p(y = 1|\text{do}(x = 0), \boldsymbol{w} = \boldsymbol{w'}) = \sigma(\beta_0(\boldsymbol{w'}))$ and $p(y = 1|\text{do}(x = 1), \boldsymbol{w} = \boldsymbol{w'}) = \sigma(\beta_0(\boldsymbol{w'}) + \beta_x)$ and β_x is given a priori by assumption.

2.2.1 Non-collapsibility

An important consideration for offset models is the difference between the *marginal* odds-ratio and the *conditional* odds-ratio. In a sufficiently large randomized controlled trial where treatment x is randomized and covariate $\boldsymbol{w}$ is observed, two different models may be estimated:

$$p(y = 1|\text{do}(x = x')) = \sigma(\gamma_0 + \gamma_x x') \tag{4}$$

$$p(y = 1|\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'}) = \sigma(\beta_0(\boldsymbol{w'})) + \beta_x(\boldsymbol{w'})x' \tag{5}$$

In contrast with linear models, even when $\beta_x(\boldsymbol{w'}) = \beta_x$ (meaning that the fixed relative treatment effect assumption holds), in general $\beta_x \neq \gamma_x$. This means that the log odds-ratio that denotes the treatment effect is different whether the model conditions on the covariate $\boldsymbol{w}$. This property is called *non-collapsibility* (Burgess, 2017). To illustrate non-collapsibility, consider the extreme example with binary w where $p(y = 1|\text{do}(x = \{0, 1\}), w = 0)) = \{0.01, 0.02\}$ and $p(y = 1|\text{do}(x = \{0, 1\}), w = 1)) = \{0.98, 0.99\}$. For both $w \in \{0, 1\}$, the *w-conditional* odds-ratio $\beta_x \approx 2.0$. However, when grouping patients with different values of w together we see that $p(y = 1|\text{do}(x = \{0, 1\}) = \{0.495, 0.505\}$, thus the *marginal* odds-ratio $\gamma_x \approx 1.0$. In the majority of randomized controlled trials the *marginal* odds ratio e^{γ_x} is estimated. If $\gamma_x \neq \beta_x$ the trials do not provide the information required to use the offset method as defined in Equation 2. As it turns out, the stronger the effect of w on the outcome, the greater the difference between γ_x and β_x becomes. This is an important drawback, as at the same time, the stronger the effect of w on the outcome, the more potential benefit the offset method has to offer for improving conditional average estimates. Later we discuss potential solutions for this issue but in our experiments we will use either β_x or γ_x directly.

2.3 Metric

The goal of the conditional average treatment effect models is to estimate the difference in outcome probability under the hypothetical interventions of treatment or no treatment. A common metric in this case is the 'Precision in Treatment Effect Heterogeneity" (PEHE, Hill (2011)). The PEHE is the root-mean-squared error of the predicted difference in outcome probability versus the actual difference in outcome probability. If $\pi_1(\boldsymbol{w}), \pi_0(\boldsymbol{w})$ denote the actual outcome probabilities under the hypothetical intervention of treatment or no treatment conditional on $\boldsymbol{w}$, and $\hat{\pi}_1(\boldsymbol{w}), \hat{\pi}_0(\boldsymbol{w})$ the predicted probabilities, the PEHE is defined as:

$$\text{PEHE} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left((\pi_1(\boldsymbol{w}_i) - \pi_0(\boldsymbol{w}_i)) - (\hat{\pi}_1(\boldsymbol{w}_i) - \hat{\pi}_0(\boldsymbol{w}_i))\right)^2}$$

3 Related Work

Estimating individual treatment effects from observational data requires assumptions. The assumption of unconfoundedness enables treatment effect estimation from observational data but this assumption is often not tenable in applications. When unconfoundedness does not hold, potential assumptions that allow for treatment effect estimation are the presence of proxy measurements of unobserved confounders (Kuroki and Pearl, 2014; Miao et al., 2016, 2018; Lee and Bareinboim, 2021; Kallus et al., 2021; van Amsterdam et al., 2021), or exploiting instrumental variables (Wald, 1940; Amemiya, 1974; Darolles et al., 2011; Hartford et al., 2017; Puli and Ranganath, 2020). These methods have been extended to neural network architectures e.g. in Kallus et al. (2021); Hartford et al. (2017); Puli and Ranganath (2020).

There are medical examples of the presented assumption of a fixed relative treatment effect for linear logistic models or survival models. Medical examples are based on breast cancer (PREDICT v2.0, Candido dos Reis et al. (2017), Adjuvant! Ravdin et al. (2001) and Adjutorium Alaa et al. (2021)), or cardiovascular disease (Xu et al., 2021). Our contribution is that we investigate the validity of the assumption that a fixed treatment effect allows for treatment effect estimation in the presence of unobserved confounding, and that we extend this assumption to the model class of neural networks.

4 Theory & Experiments

We evaluate the validity of estimating CATE models from observational data using offset models in the presence of unobserved confounding. First, we study a simple example where the expected log-likelihood is available in closed form. In this example we find that in the presence of confounding, offset models do not estimate the ground truth interventional distribution. We then study the magnitude of the resulting bias for a wide range of confounding magnitudes and find that the bias is small in many cases. As the baseline situation for CATE models is using a single average treatment effect for all patients, we study in what situations offset models can still have better PEHE than the baseline, despite the bias in the offset model. Finally, we test the offset method on a task with real-world medical images and simulated outcome data and find that it outperforms the baseline and other competing methods.

4.1 Binary confounder

A simple example compatible with Figure 2 and Equation 2 is where u is binary and $\beta_0(\boldsymbol{w}) = \beta_0^*$ for all $\boldsymbol{w}$. Denoting $\mathcal{B}$ as the Bernoulli distribution, $p_u = p(u = 1)$ and $\pi_{xu} = p(y = 1|x, u)$, then the data-generating mechanism for this example is:

$$u \sim \mathcal{B}(p_u), x \sim \mathcal{B}(p(x = 1|u = u)), y \sim \mathcal{B}(\pi_{xu}) \tag{6}$$

In the Appendix A.1 we derive a closed-form expression for the expected log-likelihood $L(\beta_0) = E_{p_{\text{obs}}(y,x,u)}[l(y, \hat{\pi}(x, u, \beta_0)]$ under the observational distribution in this example. Taking the derivative with respect to parameter β_0 and plugging in the ground truth value for β_0^* we find the following expression:

$$\frac{\partial L}{\partial \beta_0}\left(\beta_0 = \beta_0^*\right) = p_u(1 - p_u)\big[(\pi_{01} - \pi_{00})\left(p(x = 0|u = 1) - p(x = 0|u = 0)\right) + (\pi_{11} - \pi_{10})\left(p(x = 1|u = 1) - p(x = 1|u = 0)\right)\big]$$

In general this expression is non-zero, meaning that the ground truth solution β_0^* is not a stationary point of the expected log-likelihood. In the case of no confounding (i.e. $\pi_{x0} = \pi_{x1}$ or $p(x = x'|u = 1) = p(x = x'|u = 0)$) this expression is zero and β_0^* is a stationary point of the expected log-likelihood.

To evaluate the amount of bias we parameterize the magnitude of confounding using log odds-ratios $\beta_{u\to x}, \beta_{u\to y}$ so that $p(x = 1|u) = \sigma(\frac{1}{2}\beta_{u\to x}(1 - 2u))$ and $p(y = 1|x, u) = \sigma(\frac{1}{2}(\beta_x(1 - 2x) + \beta_{u\to y}(1 - 2u)))$. As there is no closed-form solution of the gradient $L(\beta_0)$ we plot the log-likelihood profile for different values of $\beta_{u\to x} = \beta_{u\to y}$ in Figure 3. Even in extreme cases of confounding when $\beta_{u\to x} = \beta_{u\to y} = \log 10$, the difference between the minimum of the expected negative log-likelihood profile in the observational setting is very close to the minimum of the expected negative log-likelihood of in the randomized trial setting where there is no confounding. This indicates that the bias induced by the unobserved confounder u is small when the assumptions hold and the offset method is used.

4.2 Binary confounder and binary covariate

The bias induced by the confounding in the simple example seems minor even for extreme magnitudes of confounding. However, the ultimate metric is whether the PEHE of the offset model is better than that of the baseline of using a single average treatment effect for each patient. To investigate this we extended the simple example by introducing a single binary covariate w with non-zero effect on the outcome. The updated data generating mechanism is:

$$u \sim \mathcal{B}(p_u), x \sim \mathcal{B}(p(x = 1|u = u)), w \sim \mathcal{B}(p_w), y \sim \mathcal{B}(\pi_{xwu}) \tag{7}$$

where

$$\pi_{xwu} = p(y = 1|x, w, u) = \sigma(\frac{1}{2}(\beta_x(1 - 2x) + \beta_w(1 - 2w) + \beta_{u\to y}(1 - 2u))) \tag{8}$$

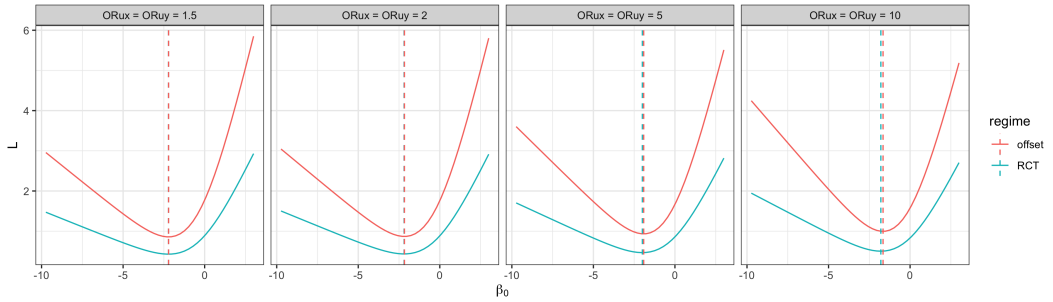

Figure 3: Expected log-likelihood profiles of β_0 for different magnitudes of confounding, indexed by $\text{OR}_{u\to x} = \text{OR}_{u\to y}$, the odds-ratios from confounder u to treatment x and outcome y respectively

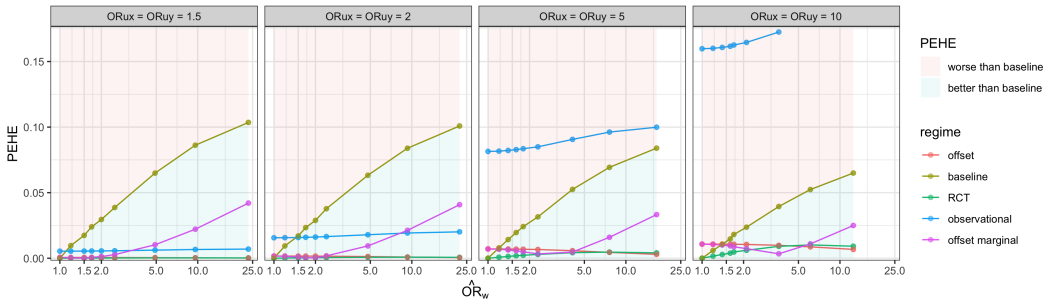

Figure 4: PEHEs for different strategies, indexed by $\text{OR}_{u\to x} = \text{OR}_{u\to y}$, the odds-ratios from confounder u to treatment x and outcome y respectively. The shaded areas indicate whether the chosen approach improves upon the baseline of assuming a single predicted difference in outcome for all patients.

4.2.1 Numerical experiments

For each value of $\beta_w, \beta_{u\to x}, \beta_{u\to y}$ as in Equation 8 the baseline PEHE is calculated. This value is contrasted with (1) the PEHE of the ideal ground truth model based on data from a randomized trial where $p_{\text{rct}}(x = 1|u = 0) = p_{\text{rct}}(x = 1|u = 1) = p_{\text{rct}}(x = 1)$, (2) a logistic regression model where both β_x, β_w are estimated from observational data, (3) an offset model where β_w is estimated while plugging in the ground truth β_x^* as obtained by the randomized trial in (1), and (4) the marginal model where the *marginal* γ_x is available from randomized controlled trials and is used as an offset in place of β_x. For these experiments we set $\beta_{u\to x} = \beta_{u\to y} = \beta_u$ to four different values and varied β_w. As seen in Figure 4, the PEHE of the observational logistic regression model becomes worse than the baseline for higher magnitudes of confounding. Also, for $\widehat{\text{OR}}_w = e^{\hat{\beta}_w} > 1$, the PEHE of the offset model is better than that of the baseline, except in the case of extreme confounding ($\beta_u = \log 10$). These results hold both when the oracle value of β_x is used or when the marginal log-odds ratio γ_x is used, though the latter performs worse when the β_w becomes stronger. However, the *marginal* offset model using γ_x is still better than the baseline whenever the oracle offset model using β_x is better than the baseline.

4.3 Lung nodules

To evaluate offset models on real-world images of lung nodules we used the open source LIDC-IDRI data-set (Armato et al., 2015). The process of creating semi-synthetic data using this data-set was described before in van Amsterdam et al. (2019). The LIDC-IDRI data-set consists of chest computed tomography (CT) scans of 1018 patients with one or more pulmonary nodules per patient. A CT-scan consists of multiple slices of images. Each nodule is visible on one or more CT-slices. Based on manual delineations from experts, a small region of 7mm around the nodule was extracted on each slide on which a nodule appeared. These segmentations are provided with the LIDC-IDRI data-set. On each slice, two measurements were taken from the nodule in the region defined by the delineation: the size of the nodule (measured in mm^2) and the heterogeneity of the intensity of the pixels in the

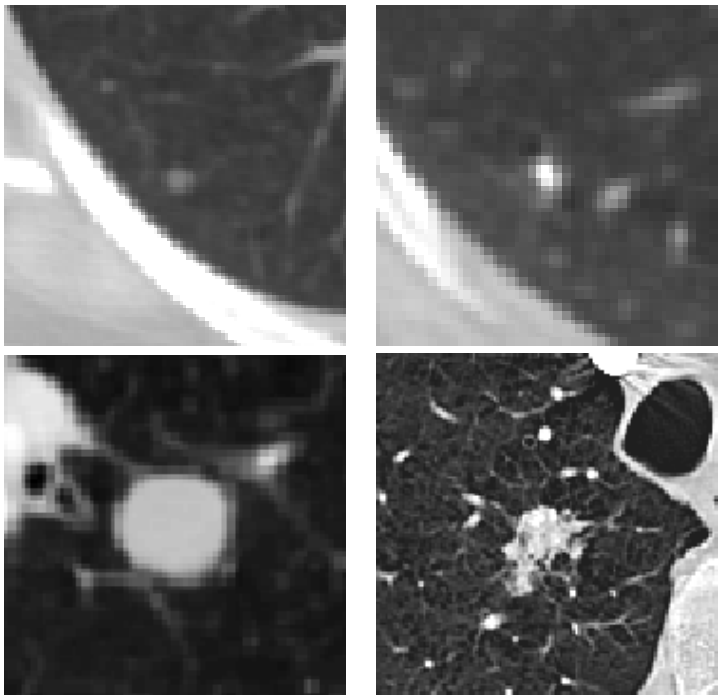

Figure 5: Four pulmonary nodules differing in size (top to bottom) and heterogeneity (left to right)

nodule measured as variance of the pixel intensities. For examples of nodules of differnet size and heterogeneity Figure 5. Both measurements were transformed using a Yeo-Johnson transformation so that their distributions resemble a standard normal distribution more closely. This results in a stack of 6568 images of lung nodules with associated measurements. The images were split randomly in a train (5000), tune (500) and test set (1068) for the experiments.

Outcome data were simulated conditional on two observed 'covariates' w_1, w_2 that represent size and heterogeneity of the nodules respectively and with a normally distributed confounder u. This ensures a statistical association between the size and heterogeneity of the nodules in the images and the simulated treatment and outcome variables. Denoting the Gaussian distribution as $\mathcal{N}$, the full data generating process is given below.

$$
\begin{aligned}
\text{image} &\sim \text{RandomUniformDraw}(\text{ImageStack}) \\
w_1, w_2 &= \text{MeasureSizeAndHeterogeneity}(\text{image}) \\
u &\sim \mathcal{N}(0, 1) \\
x &\sim \mathcal{B}(\sigma(u)) \\
y &\sim \mathcal{B}(\sigma(x + w_1 + w_2 + u))
\end{aligned}
$$

We compared three neural network methods.

As a first baseline we used TARNet (Shalit et al., 2017), which attaches two 'heads' based on a single representation, one for each treatment arm. As a second baseline we implement a convolutional neural network that instead of two separate heads has a single learnable parameter from treatment x to the log-odds of the outcome (Learnable-β_x). This variant also exploits the assumption that there is a constant difference in log-odds between the treated and untreated groups but estimates this relative treatment effect from the data. We implement the offset method using a convolutional neural network with a single output layer that uses a fixed offset depending on the treatment variable as in Equation 2. The *marginal* odds-ratio e^{γ_x} was calculated from each simulated dataset to emulate the setting where only the *marginal* treatment effect is known from a randomized controlled trial. All neural networks used the same convolutional encoder architecture for a fair comparison. Details on the

exact architectures are presented in the Appendix A.2. All experiments were repeated 10 times with different random seeds.

We report the PEHE over 10 independent realizations of simulations with an unobserved confounder in Table 1. Only Offset improves the estimation of the CATE. The other methods perform worse than the baseline of predicting a fixed absolute treatment effect for all patients.

model	baseline	PEHE	difference	sd
TARNet	0.101	0.485	0.384	0.023
Learnable-β_x	0.101	0.164	0.063	0.011
Offset	0.101	0.082	**-0.019**	0.002

Table 1: Results on image experiments in the presence of unobserved confounding, averaged over 10 different random seeds

5 Conclusion

We evaluated whether the offset method provides valid conditional average treatment effect estimates in the presence of unobserved confounding and applied this method to numerical and image data. Though not exact, the offset method may still have better PEHE than the baseline of using the average treatment effect for all patients even for large confounding magnitudes. In our numerical experiments, this holds even if an estimate of the *marginal* odds-ratio is used from randomized trials instead of the *conditional* odds-ratio.

A limitation of our work is the relatively limited set of experiments. In all our experiments $\boldsymbol{w}$ was marginally independent of x, u. Future work could experiment with different functional relationships between the variables. An important question for practical applications is when it is valid to assume that the relative treatment effect is indeed fixed, meaning that β_x does not depend on $\boldsymbol{w}$.

There may be better ways to incorporate prior knowledge from randomized trials in the form of estimates of γ_x. Given an estimate $\hat{\beta}_x$ of β_x, the implied *marginal* odds-ratio $\hat{\gamma}_x(\hat{\beta}_x)$ can be calculated as a deterministic function of $\hat{\beta}_x$ and the observed data. The marginal odds ratio estimate from a randomized trial with associated uncertainty could be used as a conditioning criterion or constraint for this derived implied marginal odds ratio. Also, future work could focus on relative treatment effect estimates in the form of hazard ratios. Finally, if there are randomized trials available where the relevant covariates $\boldsymbol{w}$ are measured but the trials are too small to estimate the entire conditional treatment effect model, a targeted maximum likelihood approach could be used to estimate the conditional odds ratio β_x while treating the $\hat{\beta}_0(\boldsymbol{w})$ function as a nuisance parameter. We leave these extensions for future work.

References

Alaa, A. M., Gurdasani, D., Harris, A. L., Rashbass, J., and van der Schaar, M. (2021). Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence*, 3(8):716–726. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Breast cancer;Prognosis Subject_term_id: breast-cancer;prognosis.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110. Publisher: Elsevier.

Armato, S., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Van Beek, E. J., Yankelevitz, D., Biancardi, A. M., Bland, P. H., Brown, M. S., Engelmann, R. M., Laderach, G. E., Max, D., Pais, R. C., Qing, D. P., Roberts, R. Y., Smith, A. R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G. W., Jude, C. M., Munden, R. F., Petkovska, I., Quint, L. E., Schwartz, L. H., Sundaram, B., Dodd, L. E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A. V., Gupte, S., Sallam, M., Heath, M. D., Kuhn, M. H., Dharaiya, E., Burns, R., Fryd, D. S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B. Y., and Clarke, L. P. (2015). Data From LIDC-IDRI. Type: dataset.

Burgess, S. (2017). Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Communications in Statistics - Theory and Methods*, 46(2):786–804. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2015.1006778.

Candido dos Reis, F. J., Wishart, G. C., Dicks, E. M., Greenberg, D., Rashbass, J., Schmidt, M. K., van den Broek, A. J., Ellis, I. O., Green, A., Rakha, E., Maishman, T., Eccles, D. M., and Pharoah, P. D. P. (2017). An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Research*, 19(1):58. 80 citations (Crossref) [2021-08-06].

Cardoso, F., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rubio, I., Zackrisson, S., and Senkus, E. (2019). Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 30(8):1194–1220.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. (2011). NONPARAMETRIC INSTRUMENTAL REGRESSION. *Econometrica*, 79(5):1541–1565. Publisher: [Wiley, Econometric Society].

Furie, R., Rovin, B. H., Houssiau, F., Malvar, A., Teng, Y. K. O., Contreras, G., Amoura, Z., Yu, X., Mok, C.-C., Santiago, M. B., Saxena, A., Green, Y., Ji, B., Kleoudis, C., Burriss, S. W., Barnett, C., and Roth, D. A. (2020). Two-Year, Randomized, Controlled Trial of Belimumab in Lupus Nephritis. *The New England Journal of Medicine*, 383(12):1117–1128.

Gradishar, W. J. (2021). NCCN Breast Cancer Guideline, Version 5.2021.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR. ISSN: 2640-3498.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Kallus, N., Mao, X., and Uehara, M. (2021). Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach. *arXiv:2103.14029 [cs, stat]*. arXiv: 2103.14029.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lean, M. E., Leslie, W. S., Barnes, A. C., Brosnahan, N., Thom, G., McCombie, L., Peters, C., Zhyzhneuskaya, S., Al-Mrabeh, A., Hollingsworth, K. G., Rodrigues, A. M., Rehackova, L., Adamson, A. J., Sniehotta, F. F., Mathers, J. C., Ross, H. M., McIlvenna, Y., Stefanetti, R., Trenell, M., Welsh, P., Kean, S., Ford, I., McConnachie, A., Sattar, N., and Taylor, R. (2018). Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. *Lancet (London, England)*, 391(10120):541–551.

Lee, S. and Bareinboim, E. (2021). Causal Identification with Matrix Equations. *Columbia CausalAI Laboratory Technical Report (R-70)*.

Miao, W., Geng, Z., and Tchetgen, E. T. (2016). Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder. *arXiv:1609.08816 [stat]*. _eprint: 1609.08816.

Miao, W., Geng, Z., and Tchetgen, E. T. (2018). Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder. *arXiv:1609.08816 [stat]*. arXiv: 1609.08816.

Puli, A. M. and Ranganath, R. (2020). General Control Functions for Causal Effect Estimation from Instrumental Variables. *arXiv:1907.03451 [cs, stat]*. arXiv: 1907.03451.

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., and Parker, H. L. (2001). Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *Journal of Clinical Oncology*, 19(4):980–991. 679 citations (Crossref) [2021-08-06].

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv:1606.03976 [cs, stat]*. arXiv: 1606.03976.

Shi, C., Blei, D. M., and Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. *arXiv:1906.02120 [cs, stat]*. arXiv: 1906.02120.

Simonovich, V. A., Burgos Pratx, L. D., Scibona, P., Beruto, M. V., Vallone, M. G., Vázquez, C., Savoy, N., Giunta, D. H., Pérez, L. G., Sánchez, M. D. L., Gamarnik, A. V., Ojeda, D. S., Santoro, D. M., Camino, P. J., Antelo, S., Rainero, K., Vidiella, G. P., Miyazaki, E. A., Cornistein, W., Trabadelo, O. A., Ross, F. M., Spotti, M., Funtowicz, G., Scordo, W. E., Losso, M. H., Ferniot, I., Pardo, P. E., Rodriguez, E., Rucci, P., Pasquali, J., Fuentes, N. A., Esperatti, M., Speroni, G. A., Nannini, E. C., Matteaccio, A., Michelangelo, H. G., Follmann, D., Lane, H. C., Belloso, W. H., and PlasmAr Study Group (2021). A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia. *The New England Journal of Medicine*, 384(7):619–629.

van Amsterdam, W. A. C., Verhoeff, J. J. C., de Jong, P. A., Leiner, T., and Eijkemans, M. J. C. (2019). Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. *npj Digital Medicine*, 2(1):1–6. Number: 1 Publisher: Nature Publishing Group.

van Amsterdam, W. A. C., Verhoeff, J. J. C., Harlianto, N. I., Bartholomeus, G. A., Puli, A. M., Jong, P. A. d., Leiner, T., Lindert, A. S. R. v., Eijkemans, M. J. C., and Ranganath, R. (2021). Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer. *medRxiv pre-print*, page 2021.10.30.21265597. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Type: article.

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. *The Annals of Mathematical Statistics*, 11(3):284–300. Publisher: Institute of Mathematical Statistics.

Watson, T. (2007). Practitioner's Guide to Generalized Linear Models. Technical report, Towers Watson.

Xu, Z., Arnold, M., Stevens, D., Kaptoge, S., Pennells, L., Sweeting, M. J., Barrett, J., Di Angelantonio, E., and Wood, A. M. (2021). Prediction of Cardiovascular Disease Risk Accounting for Future Initiation of Statin Treatment. *American Journal of Epidemiology*, page kwab031.

A Appendix

A.1 Identification

We now prove that the assumption expressed in Equation 2 is not sufficient for identifying the interventional distribution $p(y = 1|\text{do}(x = x'), \boldsymbol{w} = \boldsymbol{w'})$ from observational data using a simple example where all variables are binary and $\beta_0(\boldsymbol{w'}) = \beta_0^*$ for all $\boldsymbol{w'}$. We first derive an expression for the expected log likelihood $L(\beta_0) = E_{p_{\text{obs}}(y,x,u)}[l(y, \hat{\pi}(x, u, \beta_0)]$ under the observational distribution in this example. Writing

$$\begin{aligned} p_u =& p(u = 1) \\ p_{x'u'} =& p(x = x', u = u') = p(x = x'|u = u')p(u = u') \\ \pi_{x'u'} =& p(y = 1|x = x', u = u') \end{aligned}$$

Then the data generating mechanism is:

$$u, x \sim \mathcal{B}(p_{x'u'}), y \sim \mathcal{B}(\pi_{xu})$$

The ground truth solutions for β_0^* and β_x^* are:

$$\begin{aligned} p(y|\text{do}(x = 0)) &= (1 - p_u)\pi_{00} + p_u\pi_{01} = \sigma(\beta_0^*) \\ p(y|\text{do}(x = 1)) &= (1 - p_u)\pi_{10} + p_u\pi_{11} = \sigma(\beta_0^* + \beta_x^*) \end{aligned}$$

The Bernoulli log-likelihood is

$$l(y|x, \beta_0, \beta_x) = y \log \sigma(\beta_0 + \beta_x x) + (1 - y) \log(1 - \sigma(\beta_0 + \beta_x))$$

Here β_x is assumed fixed and β_0 is the only parameter, resulting in the following expression for $L(\beta_0)$:

$$\begin{aligned} L(\beta_0) =& p_{00} \left[\pi_{00} \log \sigma(\beta_0) + (1 - \pi_{00}) \log(1 - \sigma(\beta_0))\right] \\ &+ p_{01} \left[\pi_{01} \log \sigma(\beta_0) + (1 - \pi_{01}) \log(1 - \sigma(\beta_0))\right] \\ &+ p_{10} \left[\pi_{10} \log \sigma(\beta_0 + \beta_x) + (1 - \pi_{10}) \log(1 - \sigma(\beta_0 + \beta_x))\right] \\ &+ p_{11} \left[\pi_{11} \log \sigma(\beta_0 + \beta_x) + (1 - \pi_{11}) \log(1 - \sigma(\beta_0 + \beta_x))\right] \end{aligned}$$

Taking the derivative with respect to β_0 and noting that $(\log \sigma(x))' = 1 - \sigma(x)$ we arrive at

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} =& p_{00} \left[\pi_{00}(1 - \sigma(\beta_0)) - (1 - \pi_{00})\sigma(\beta_0)\right] \\ &+ p_{01} \left[\pi_{01}(1 - \sigma(\beta_0)) - (1 - \pi_{01})\sigma(\beta_0)\right] \\ &+ p_{10} \left[\pi_{10}(1 - \sigma(\beta_0 + \beta_x)) - (1 - \pi_{10})\sigma(\beta_0 + \beta_x)\right] \\ &+ p_{11} \left[\pi_{11}(1 - \sigma(\beta_0 + \beta_x)) - (1 - \pi_{11})\sigma(\beta_0 + \beta_x)\right] \end{aligned}$$

Plugging in the ground truth solutions β_0^*, β_x^* and re-arranging we arrive at:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0}\left(\beta_0 = \beta_0^*, \beta_x = \beta_x^*\right) = p_u(1 - p_u)\big[&(\pi_{01} - \pi_{00})\left(p(x = 0|u = 1) - p(x = 0|u = 0)\right) + \\ &(\pi_{11} - \pi_{10})\left(p(x = 1|u = 1) - p(x = 1|u = 0)\right)\big] \end{aligned}$$

If there is no confounding this expression is zero, but in general it is not which means that the ground truth solution β_0^* is not an optimum of the expected log-likelihood in the observational data distribution.

Of note, the fact that the interventional distribution is not identified does not automatically imply that the CATE is not identified as there may be another $\beta_0' \neq \beta_0^*$ such that $\text{CATE}(\beta_0 = \beta_0', \beta_x = \beta_x^*) = \text{CATE}(\beta_0 = \beta_0^*, \beta_x = \beta_x^*)$. To investigate this, assume that for some $\beta_0^* = a$ and $\beta_x^* = b$ we have that:

$$
\begin{aligned}
\delta :=& \text{CATE}(\beta_0 = a, \beta_x = b) \\
=& \sigma(a+b) - \sigma(b) \\
=& \frac{e^{a+b}}{1+e^{a+b}} - \frac{e^a}{1+e^a}
\end{aligned}
$$

We will now prove that this equation has at most two solutions by noting that:

$$
\begin{aligned}
\frac{e^{a+b}}{1+e^{a+b}} - \frac{e^a}{1+e^a} &= \frac{e^{a+b}(1+e^a) - (1+e^{a+b})e^a}{(1+e^{a+b})(1+e^a)} \\
&= \frac{e^a(e^b-1)}{(1+e^{a+b})(1+e^a)}
\end{aligned}
$$

Introducing $y := e^a$ and cross-multiplying we get:

$$
\begin{aligned}
\delta = \frac{y(e^b-1)}{(1+e^b y)(1+y)} &\iff \\
\delta(1+e^b y)(1+y) = y(e^b-1) &= \\
\delta + \delta(1+e^b)y + \delta e^b y^2 = y(e^b-1) &\iff \\
\delta e^b y^2 + \Big(e^b(\delta-1) + \delta + 1\Big) y + \delta = 0 &
\end{aligned}
$$

Depending on the values of δ and b this quadratic equation in y has 0, 1 or 2 real-valued solutions, yielding 0, 1 or 2 real-valued solutions for $a = \log y$. This implies that there exists utmost one alternative solution $\beta_0' \neq \beta_0^*$ such that $\text{CATE}(\beta_0 = \beta_0', \beta_x = \beta_x^*) = \text{CATE}(\beta_0 = \beta_0^*, \beta_x = \beta_x^*)$. Given the results of the numerical experiments it is highly unlikely that this coincides with the optimum of the observational expected log-likelihood.

A.2 Neural Network Architectures

The convolutional encoder consisted of 5 layers of 3x3 convolutions with 16 feature channels, followed by a rectified linear unit activation function and 2x2 max-pooling. The result of this encoding was flattened into a 144-dimensional feature vector that was passed on to the final layer. These decisions were based on earlier work in van Amsterdam et al. (2019) using the same dataset. This encoder architecture was used for all models. For Offset and Learnable-β_x, the output layer consisted of a single linear layer with a 1D output. For Offset a fixed offset was added to the predicted log-odds based on the treatment. For Learnable-β_x this treatment effect was a learnable scalar parameter. For TARNet, two linear layers with a single 1D output each were added. Depending on the value of x one of both heads provided the final prediction. All models yielded log-odds predictions under hypothetical treatment or no-treatment. In Learnable-β_x, the treatment effect parameter was initialized to a value of 1. In TARNet, the bias of the output of one head was initialized to 0 and the bias of the output of the other head was initialized to 1. All other parameters were randomly initialized with the default initialization scheme in PyTorch. PyTorch version 1.7 was used. We used the Adam optimizer with a learning rate of 0.005 and default hyperparameters, the batch size was 200. The total time for all experiments was under 8 hours on a single NVIDIA P-6000 GPU.

A.3 Image Data and Simulations

The LIDC-IDRI data-set is released on under a Creative Commons Attribution 3.0 Unported License. The data from LIDC-IDRI stem from seven hospitals. The study was approved by the appropriate Institutional Review Boards, including the informed consent procedure. The scans were anonymized by removing identifiable metadata.