Toward an Understanding of Adversarial Examples in Clinical Trials: Supplementary Material

Konstantinos Papangelou[1], Konstantinos Sechidis[1], James Weatherall[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.papangelou, konstantinos.sechidis, gavin.brown}@manchester.ac.uk
[2] Advanced Analytics Centre, Global Medicines Development, AstraZeneca,Cambridge, SG8 6EE, UK
James.Weatherall@astrazeneca.com

1 Common Assumptions in Causal Inference

For treatment effects to be identifiable from the observed data certain assumptions have been identified in the literature [8, 10] and have been adopted for the estimation of ITE in observational studies (e.g. see [11] for a recent example).

Assumption 1. *(Stable Unit Treatment Value Assumption - SUTVA) There is a single version of each treatment and there is no interference between the subjects.*

This assumption guarantees that all subjects have access to the same treatment and a treatment applied to one subject does not affect the outcome for another subject. This assumption (along with consistency, which describes that for each subject we only observe the outcome under the actual treatment received) allows us to express the (counter)factual outcome as a function of the potential outcomes and the treatment.

Assumption 2. *(Unconfoundedness) For a subject $\boldsymbol{X}$, the treatment assignment is statistically independent of the potential outcomes: $(Y_1, Y_0) \perp T \mid \boldsymbol{X}$.*

The unconfoundedness assumption (also know as *no-hidden confounders*, *strong ignorability*, *selection on observables* [6]) implies that among those subjects that have the same characteristics $\mathbf{X}$, the treatment assignment is random. In RCTs the assignment mechanism is known and depends only on observed features and therefore unconfoundedness holds by design. In observational studies, this assumption cannot be ensured since we do not have knowledge of the treatment assignment mechanism and it cannot always be verified by the data. Its validity must be determined by understanding the causal relationships between the features, the outcome and the treatment assignment.

Assumption 3. *(Overlap) For a subject $\boldsymbol{X} = \boldsymbol{x}$, the probability of receiving each one of the possible treatments is bounded away from zero:* $0 < p(T = 1 \mid \boldsymbol{x}) < 1$.

The last assumption (also referred to as *common support*, *positivity*, *balance*) guarantees that there is enough randomness in the treatment assignment procedure so that each subject has a non-zero probability of receiving one of the possible treatments. In terms of the observed data, this assumption guarantees that there will be enough overlap in the feature space, so that for each subject that received treatment $T = 1$, there is at least one similar subject that received the opposite treatment. The last two assumptions are usually referred to as *strong ignorability* [8]. In this work, we focused on randomised experiments where strong ignorability holds by design.

2 Comments on the Plausibility of Adversarial Examples and Discrete Spaces

Initial works on adversarial examples showed that they act as blind spots for the model, raising questions however about their plausibility. For example, MNIST images are essentially binary and as a result creating adversarial examples that modify white and black pixels slightly will result in images that are misclassified by state-of-the-art models while they might look similar [4]. However, to what extend these images might occur in the real world is questionable. In the experiments performed on IHDP, we rounded each feature to the closest value it can take based on its domain in order to identify adversarial patients that may occur in the real world. Note that the resulting adversarial patients can be biologically plausible, but the chances of occurring them in the real world needs to be verified by domain experts. Without this post-processing step, the accuracy on the factual outcome would have dropped rapidly even with smaller values of θ. The resulting adversarial patients in this case could be considered as blind spots for the model, but their practical implications would be limited since they will never occur in practice. Still the effect of these (non-realistic) adversarial examples could be considered as a regulariser through adversarial training.

We can identify adversarial patients in discrete spaces by selecting the closest patient that best aligns with the adversarial direction as described by Biggio et al. [2]. In this case, in order to create realistic adversarial patients, more post-processing steps are required in order to ensure that each feature takes a value within its domain, but also that its value does not contradict with semantically similar features (e.g. a person cannot have both one child and twins).

3 Experimental Protocol

For IHDP we used the response surface B of Hill [5] to derive the continuous outcomes and then we formed a binary classification task so that the treatment has a constant average effect $ATE = 0.2$ (on the probability space). To select the hyperparameters of the network for IHDP we performed random search within

Table 1. Parameters search space

Parameter	Range
Representation layers	$\{1, 2, 3\}$
Task-specific layers	$\{1, 2, 3\}$
Representation layer units	$\{10, 20, 50, 100, 200\}$
Task-specific layer units	$\{10, 20, 50, 100, 200\}$
Dropout probability	$\{0, 0.2, 0.5\}$
Weight decay	$\{0, 1e^{-4}, 1e^{-3}\}$
Learning rate	$\{1e^{-4}, 1e^{-3}\}$

the search space shown in table 1 and selected the setting that achieved the lowest factual loss on the validation set. For validation the results have been averaged over 100 realisations of the outcome for IHDP and 50 for the synthetic data. The results on the test set have been averaged over 1000 realisations of the outcome for IHDP and 100 for the synthetic data. To avoid overfitting we applied early stopping using the validation loss. For IHDP we trained a network with 2 shared layers and 2 group-specific. The size of the layers were 50. For the simulated models we used smaller networks with up to 3 layers. In all networks we used ReLU as the activation function and applied dropout with probability 0.5. The models were trained using Adam [7].

The adversarial examples were created using an iterative approach with $m = 10$ iterations and a constant step $\alpha = \theta/m$, where θ is the total perturbation. In order to perform adversarial training efficiently, we followed a single-shot approach. In our experiments we observed that adversarial training acts as a strong regulariser especially if used in combination with dropout. To achieve fair comparisons we used the same architectures and the same procedure for early stopping. We observed that adversarial training allows us to train for more epochs, before starting to overfit. These observations are in accordance with recent results in adversarial training [4, 1].

Table 2. Simulated outcome functions

SM1	$\text{logit}(f(\mathbf{x}, T)) = -1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.5X_2X_7 +$ $+2T\mathbb{1}(\mathbf{x} \in \{X_1 < 0.545 \cap X_2 > -0.545\})$
SM2	$\text{logit}(f(\mathbf{x}, T)) = -1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.5X_2X_7 +$ $+2T\mathbb{1}(\mathbf{x} \in \{X_3 < 0.545 \cap X_4 > -0.545\})$
SM3	Same as SM1 but odd(even) features have an internal correlation $\rho = 0.7$
SM4	$\text{logit}(f(\mathbf{x}, T)) = -1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.5X_2X_7 +$ $0.1T + T\mathbb{1}(\mathbf{x} \in \{X_1 < 0 \cap X_2 > 0\})$

4 Results in Subgroup Identification

In this section we present additional results on SM1, used in the main paper, as well as results on SM3, SM4 (table 2). In all modifications we observed a similar trend with the results presented in the main paper.

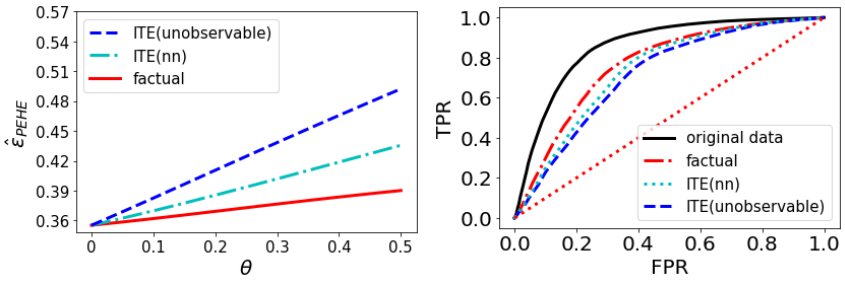


Fig. 1. We validate the existence of adversarial patients on SM3. We created adversarial patients modifying only prognostic/irrelevant features. We report the effect of adversarial patients on subgroup identification for $\theta = 0.4$.

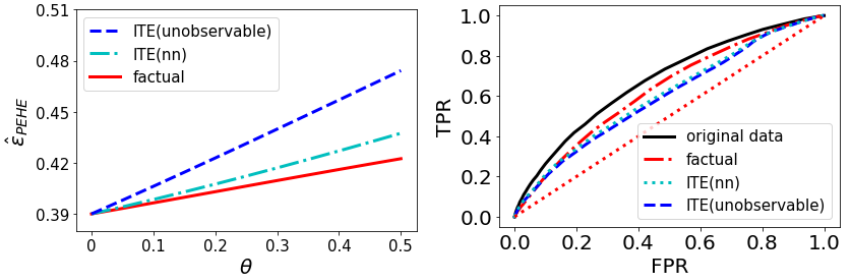


Fig. 2. We validate the existence of adversarial patients on SM4. We created adversarial patients modifying only prognostic/irrelevant features and we report the effect of adversarial patients on subgroup identification for $\theta = 0.4$.

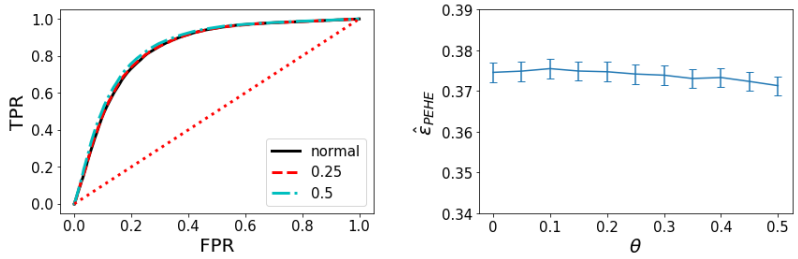


Fig. 3. To see whether any random perturbation of the non-predictive part can have an effect, we performed noisy training on SM1 by adding in the batch randomly perturbed examples with the same predictive features. This did not have any effect on the generalisation performance.

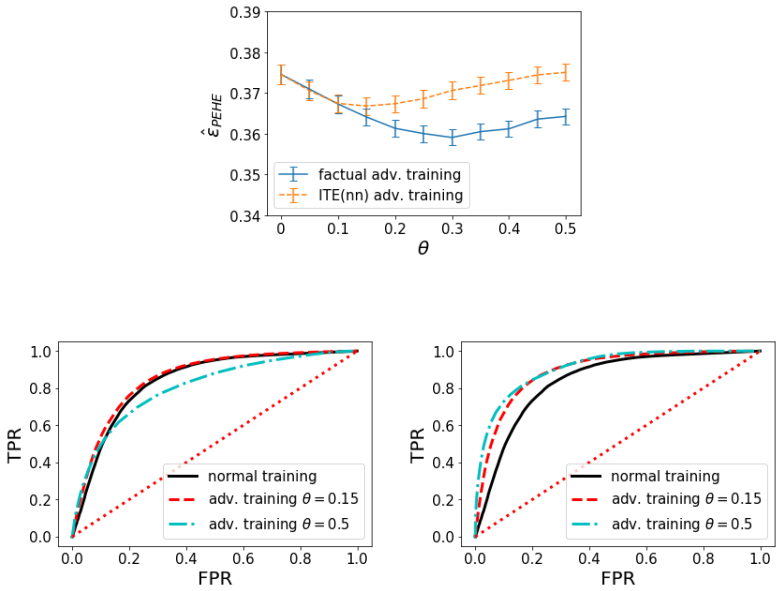
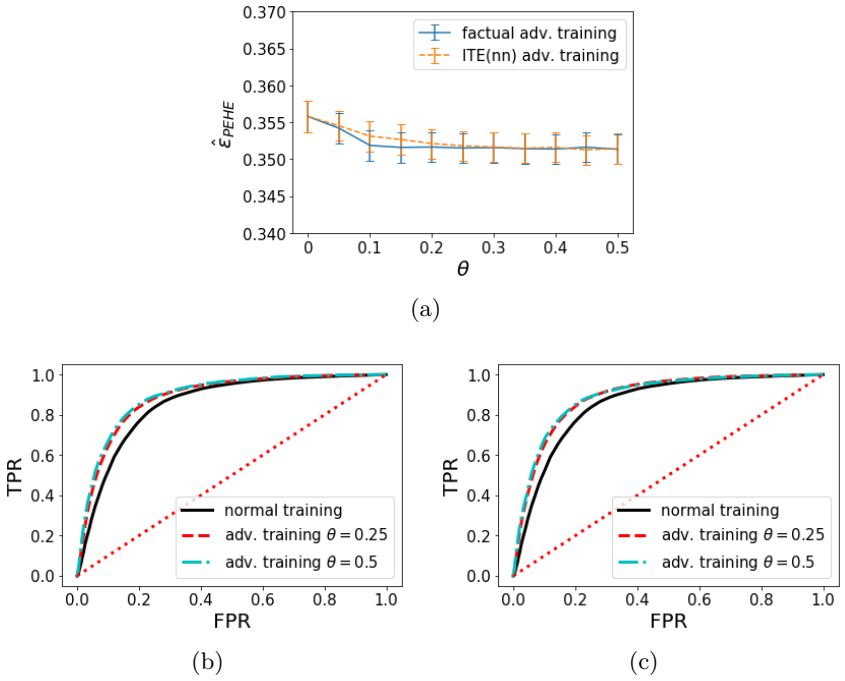


Fig. 4. We repeated the experiment described in sect. 5.2 for SM1 but assuming that we do not have any information about which features have the strongest predictive effect. Notice that for small values of θ adversarial training with respect to the factual outcome acts as a stronger regulariser and leads to improved results. Even in this extreme case that we do not have any indication about potentially prognostic/predictive features, adversarial training can be beneficial. Notice that this is unlikely to occur in practice and it is common to know at least a subset of the prognostic features, as they are used for randomisation (among other tasks) [9].



(a)

(b) (c)

Fig. 5. We modified the outcome function used in sect. 5.2, so that even (odd) numbered features have an internal correlation of 0.7 (SM3). We report the results performing adversarial training with respect to ITE using the nearest neighbour approximation to craft the adversarial patients (graph (c)) as well as the factual outcome (graph (d)). Here adversarial training with respect to the factual outcome performs marginally better for small values of θ.

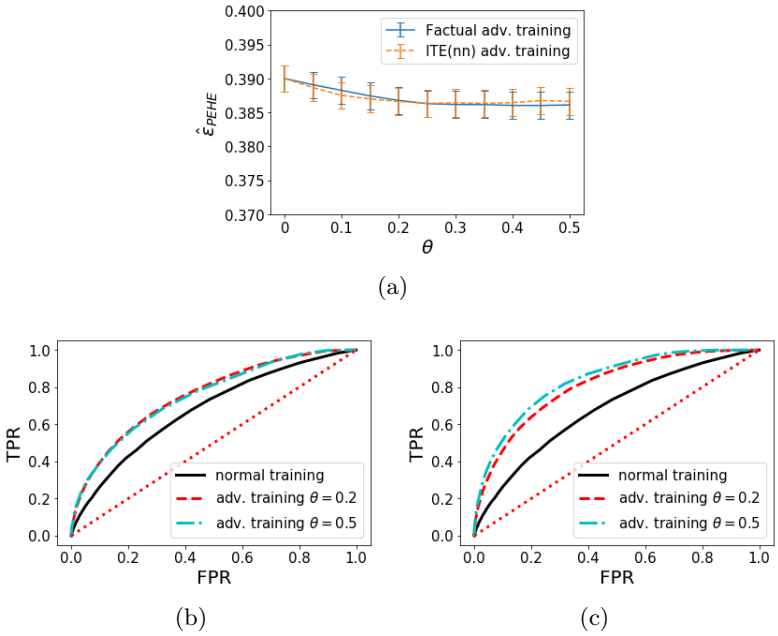


Fig. 6. We modified the outcome function used in Sect. 5.2, so that the subgroup has smaller size, but also smaller predictive strength (SM4). We report the results performing adversarial training with respect to ITE using the nearest neighbour approximation to craft the adversarial patients (graph (c)) as well as the factual outcome (graph (d)). Notice that this is a challenging scenario since the subgroup is 25% of the observations and there is a small predictive strength ($\lambda = 1$).

References

1. Arpit, D., Jastrzbski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj T., Fischer A., Courville A., Bengio Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. International Conference on Machine Learning, in PMLR 70:233-242 (2017)
2. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Roli, F.: Evasion attacks against machine learning at test time. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 387-402. Springer, Berlin, Heidelberg (2013)
3. Foster, J. C., Taylor, J. M., Ruberg, S. J.: Subgroup identification from randomized clinical trial data. Statistics in medicine, 30(24), 2867-2880 (2011)
4. Goodfellow, I. J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. International Conference on Learning Representations (2015)
5. Hill, J. L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1), 217-240 (2011)
6. Imbens, G. W., Wooldridge, J. M.: Recent developments in the econometrics of program evaluation. Journal of economic literature, 47(1), 5-86 (2009)
7. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
8. Rosenbaum, P. R., Rubin, D. B.: The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55 (1983)
9. Ruberg, S. J., Shen, L.: Personalized medicine: Four perspectives of tailored medicine. Statistics in Biopharmaceutical Research, 7(3), 214-229 (2015)
10. Rubin, D. B.: Bayesian inference for causal effects: The role of randomization. The Annals of statistics, 34-58 (1978)
11. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. Proceedings of the 34th International Conference on Machine Learning, in PMLR 70:3076-3085 (2017)
12. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)