Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou[1][0000−0001−5127−3170], Konstantinos Sechidis[1][0000−0001−6582−7453], 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

Abstract. Deep learning systems can be fooled by small, worst-case perturbations of their inputs, known as adversarial examples. This has been almost exclusively studied in supervised learning, on vision tasks. However, adversarial examples in *counterfactual* modelling, which sits outside the traditional supervised scenario, is an overlooked challenge. We introduce the concept of *adversarial patients*, in the context of counterfactual models for clinical trials—this turns out to introduce several new dimensions to the literature. We describe how there exist multiple *types* of adversarial example—and demonstrate different consequences, e.g. ethical, when they arise. The study of adversarial examples in this area is rich in challenges for accountability and trustworthiness in ML–we highlight future directions that may be of interest to the community.

Keywords: Adversarial Examples · Counterfactual Modelling · Randomised Clinical Trials · Subgroup Identification

1 Introduction

For personalised medicine, a major goal is to predict whether or not a patient will benefit from a particular treatment. The challenge here is to model the outcome of a patient under different treatment scenarios. This task sits outside traditional supervised learning, phrased as a causal inference problem, i.e. modelling the causal effect of a treatment on the outcome of a patient. Recently, Deep Neural Networks (DNNs) have been adopted in this area, achieving significant improvements on the estimation of individual-level treatment effects [1, 8, 17]. While DNNs have proven their merits in various domains, it has also been shown that they are susceptible to *adversarial examples*—small perturbations of the data, carefully designed to deceive the model. This area has received significant attention from the community, e.g. [3, 6, 13, 18]. When DNNs are used in safety-critical applications, such as healthcare, *accountability* becomes crucial [4]. One such application, where the errors may result in personal, ethical, financial and legal consequences, is in personalised medicine.

While adversarial examples have been studied extensively in the traditional supervised setting, their properties in counterfactual models create significant new challenges. We introduce the concept of *adversarial patients*, the analogue of adversarial examples in the counterfactual (healthcare) scenario. We show that in counterfactual models there exist adversarial directions that have not been examined before. By extending well-established results from supervised learning to the counterfactual setting, we show how the derived *adversarial patients* may affect clinical decisions. We note that, in the supervised adversarial literature, a common fear is the creation of *intentional* adversarial examples, creating security risks. In contrast for healthcare, the caution should be for *unintentional* adversarial patients, with edge-case health characteristics, leading to significant ethical dilemmas. Drug development is a time consuming and expensive process spanning sometimes 10-15 years [14], where the end result may affect the lives of millions, thus we argue that the study of adversarial patients is important and worthy of investigation. In particular, we note:

- They can act as warning flags—cases where the deployed model may go wrong, increasing the level of trust.
- They are in accordance with recent regulations for interpretable and accountable procedures in ML [4].

In Sect. 2 we review the concept of adversarial examples in traditional supervised problems, and in Sect. 3 the basics of counterfactual modelling, which sets our focus. From Sect. 4 onward, we introduce the concept of *adversarial patients* in counterfactual models. We present an empirical study in Sect. 5, and conclude with a discussion of promising future directions.

2 Background: Adversarial Machine Learning

Adversarial examples are carefully crafted inputs that deceive ML models to misclassify them [18]. In vision tasks such adversarial images are indistinguishable (to the human eye) from their original counterparts, but can fool a classifier into making incorrect predictions with high confidence [6, 13, 18, 12, 9]

To formally define adversarial examples, we denote a single example as the tuple $\{\mathbf{x}, y\}$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{0, 1\}$, and the class prediction from some model $\hat{f}$ is given by $h(\mathbf{x}) = \mathbb{1}(\hat{f}(\mathbf{x}) > 0.5)$. Suppose now we have a correct classification, i.e. $h(\mathbf{x}) = y$. An adversarial example, $\mathbf{x}_{adv}$, can be constructed solving the following optimisation problem, where the particular l-norm is chosen per application:

$$\underset{\mathbf{x}_{adv}}{\arg\min} \quad \|\mathbf{x} - \mathbf{x}_{adv}\|_l \qquad \text{s.t.} \quad h(\mathbf{x}_{adv}) \neq y \tag{1}$$

i.e. the closest input vector which the model $\hat{f}$ misclassifies. Note that this requires knowledge of a ground truth y, which, as we will see, is not the case in counterfactual models. Different choices of the distance measure and the optimisation procedure have resulted in many variations (e.g. [6, 13, 12, 9]). Among

those, a well-established approach, which we will adopt in this work, is the *fast gradient sign method* (FGSM) [6], which fixes the maximum perturbation and maximises an arbitrary, differentiable loss function $J(\cdot)$, as so:

$$\arg\max_{\mathbf{x}_{adv}} \quad J(\hat{f}(\mathbf{x}_{adv}), y) \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{x}_{adv}\|_\infty \leq \theta, \quad \text{for some} \quad \theta > 0 \tag{2}$$

Once adversarial examples are created, they can deceive the model $\hat{f}$ in deployment. A complementary field of study has emerged, in defending models against such examples. A common method is *adversarial training*, where they are included in a new round of training, iteratively increasing the robustness of the model. Research has shown adversarial training to be an effective regulariser [6, 18, 11], especially in combination with others for DNNs, such as Dropout.

The study of adversarial examples is especially important in the healthcare domain, where incorrect predictions may have grave repercussions. However, the standard supervised learning paradigm does not always apply in healthcare. In the following section we introduce the basics of *counterfactual modelling*, which brings new dimensions to the adversarial literature.

3 Counterfactual Modelling in Clinical Trials

3.1 The Potential Outcomes Framework

A clinical trial is a time consuming and expensive procedure with several phases. At each phase more participants are recruited in order to evaluate the safety and efficacy of a drug before it is deployed. In phases I and II preliminary results on the safety of the drug are gathered, normally using a few hundred patients. Phases III and IV are of particular relevance to the ML community, since predictive models are usually built to estimate properties of the drug. In phase III an essential task is the identification of subgroups of patients with good response to the drug, usually requiring a model to estimate individual level treatment effects as an intermediate step [5, 10]. In phase IV the drug is used in clinical practice, and a predictive model (or rules derived from it) might assist clinicians to estimate the likely outcome for a particular patient.

Inference of individual treatment effects can be formulated as a causal inference problem—for each patient we have some baseline (pre-treatment) features, the administered treatment, and the outcome. In contrast to supervised problems, we do not observe the outcome under the *alternative treatment*, i.e. we do not observe the *counterfactual* outcome. We can address this class of problems with the *potential outcomes framework* [15].

We define the treatment variable T, as a random variable[1] that indicates the treatment received by each subject; here for simplicity we assume $T \in \{0, 1\}$. The potential outcomes under the two treatment scenarios are functions of the subjects $\mathbf{X}$, denoted as $Y_1(\mathbf{X}), Y_0(\mathbf{X})$, while the observed outcome is $Y(\mathbf{X})$. For simplicity, we will omit the dependence of the outcomes on $\mathbf{X}$. For each subject

[1] We use upper-case for random variables and lower-case for their realisations.

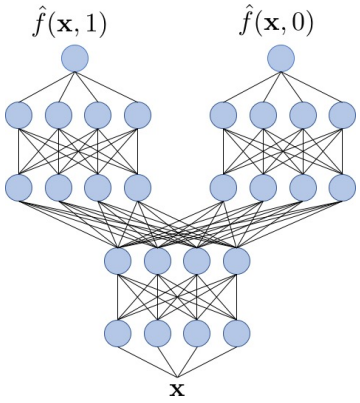


Fig. 1. Example of a multi-task counterfactual neural network.

we only observe the outcome under the actual treatment received: if $T = 0$ then $Y = Y_0$, and if $T = 1$ we observe $Y = Y_1$. The difficulty arises from the fact for any given patient, we cannot observe both Y_0 and Y_1.

The subjects that receive the drug ($T = 1$) are referred to as the *treated group*, while those who receive the baseline treatment ($T = 0$) are the *control group*. For subjects $\mathbf{X}$ the observed factual outcome can be expressed as a random variable in terms of the two potential outcomes: $Y = TY_1 + (1 - T)Y_0$. Given a set of subjects with their assigned treatments and their factual outcomes, the goal of *counterfactual modelling* is to train a model that will allow us to infer the counterfactual outcome that would have occurred if we had flipped the treatment: $Y^{cf} = (1 - T)Y_1 + TY_0$.

Building on this, the *individual treatment effect* (ITE) of a subject $\mathbf{X} = \mathbf{x}$ can be expressed as the expected difference between the two potential outcomes:

$$ITE(\mathbf{x}) := \mathbb{E}_{Y_1 \sim p(Y_1|\mathbf{x})}[Y_1 \mid \mathbf{x}] - \mathbb{E}_{Y_0 \sim p(Y_0|\mathbf{x})}[Y_0 \mid \mathbf{x}]$$

For treatment effects to be identifiable from the observed data, certain assumptions must hold (Sect. 1 of the supplementary material) and have been adopted for the estimation of ITE in observational studies (e.g. [17] for a recent example). We will focus on the case of randomised clinical trials, where *strongly ignorable* treatment assignment (i.e. treatment assignment depends only on the observed features—in our case is random) holds by design.

3.2 NNs for Estimation of Individual Treatment Effects

The adaptation of supervised learning models for causal inference tasks has gained much recent attention, both for estimation of individual treatment effects [2, 20] and for subgroup identification [5, 10]. Multi-task Deep Neural Nets have shown significant improvement over traditional baselines [1, 17], capturing the interactions between the two groups in a shared representation, while learning

the response surface for each group separately. In multi-task models, learning can be achieved following an "alternating" approach during training, where at each iteration we use either a treated or a control batch [1]. At every iteration we update the weights of the shared layers and, depending on the batch, only the corresponding task-specific layers, i.e. for a treatment batch, only the left branch in Fig. 1 is updated, and the parameters of the right branch are fixed.

Note that in contrast to Sect. 2, now we will have two outputs, one for each treatment group. These can be seen as separate models. Let us denote each output as $\hat{f}(\mathbf{x}, t), t \in \{0, 1\}$. The individual treatment effect can be estimated from the outputs of the two branches as, $\hat{\text{ITE}}(\mathbf{x}) = \hat{f}(\mathbf{x}, 1) - \hat{f}(\mathbf{x}, 0)$. A common evaluation measure of ITE that has been used extensively in the literature [1, 8, 17] is the *expected Precision in Estimation of Heterogeneous Effect* (PEHE):

$$\varepsilon_{PEHE} = \int ((\hat{f}(\mathbf{x}, 1) - \hat{f}(\mathbf{x}, 0)) - ITE(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} \tag{3}$$

The use of the true ITE assumes the knowledge of the ground truth for both potential outcomes and therefore can be estimated only in (semi-)synthetic scenarios. In real-world applications, a common approximation of ε_{PEHE} is to estimate it using the factual outcome of the nearest neighbour that received the opposite treatment [8, 17].

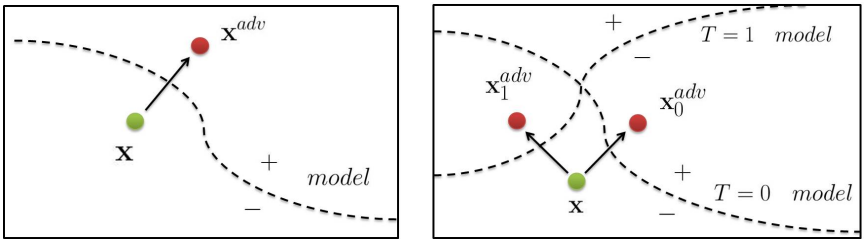
Counterfactual modelling differs significantly from traditional supervised settings, and introduces new challenges that have not been examined in the adversarial learning literature. The key contributions of this paper are:

- We show that counterfactual models introduce new sets of adversarial directions, some of which cannot be observed, but provably still exist.
- We show how we can use FGSM to identify adversarial examples with respect to the observed (factual) outcome and discuss their implications.
- We highlight that for subgroup identification (common in clinical trials) we may not be interested in the outcomes, but the *individual treatment effect*. We show that the ITE adds a new set of adversarial directions, the implications of which we shall discuss.
- We demonstrate the effect of adversarial training on the generalisation performance of counterfactual models.
- Building on our analysis, we present several new research directions, which may be of interest to the community (Sect. 6).

4 A Theoretical Study of Adversarial Directions

In this section we introduce the concept of adversarial patients in counterfactual models. We will assume a *hypothetical* scenario, where we know both potential outcomes. This will allow us to show that in counterfactual models there can exist multiple adversarial directions, with potentially different consequences.

From now on, we will focus on the case of binary outcomes and binary treatment. We will denote a single example as $\{\mathbf{x}, y_1, y_0\}$, where $\mathbf{x} \in \mathbb{R}^d$ and



(a) Traditional adversarial setting (b) Counterfactual adversarial setting

Fig. 2. In traditional adversarial settings there is a single model which we are trying to fool [3, 6]. Counterfactual adversarial examples, or "adversarial patients" can exist in several forms. Here we illustrate how the adversarial patient that fools the treatment model may be orthogonal to that which fools the control model. Note that in practice, since we know one of the outcomes, we can identify only one of the two directions.

$y_1, y_0 \in \{0, 1\}$ are the potential outcomes of $\mathbf{x}$. Note that in practice for a given $\mathbf{x}$ we observe only one of them. As we already mentioned, we will use $\hat{f}(\mathbf{x}, t)$ to denote the output of the counterfactual network. For binary outcomes, this is the probability of observing a positive outcome with $(t = 1)$ or without treatment $(t = 0)$, estimated by a model $\hat{f}(\cdot, t)$. The predicted outcomes of a model will be denoted as $h(\mathbf{x}, t) = \mathbb{1}(\hat{f}(\mathbf{x}, t) > 0.5)$.

4.1 Adversarial Patients and Potential Outcomes

Let us assume for now the hypothetical scenario, in which both potential outcomes are observable. In traditional supervised problems there will be a single model $\hat{f}$, which we are trying to deceive. On the other hand, in counterfactual modelling for two levels of treatment there will be two models, as shown in Fig. 2. The model $\hat{f}(\cdot, 1)$ defines the decision boundary between the patients who have outcome $y_1 = 1$ and $y_1 = 0$ while the model $\hat{f}(\cdot, 0)$ distinguishes between those who have $y_0 = 1$ and $y_0 = 0$. This results in the following definition of counterfactual adversarial examples, or *adversarial patients*.

Definition 1 (Adversarial Patient and Potential Outcomes). *An adversarial patient with respect to its potential outcome* y_t, $\mathbf{x}_t^{adv} \in \mathbb{R}^d$, *where* $t \in \{0, 1\}$ *is a solution to the following optimisation problem:*

$$\arg\min_{\mathbf{x}_t^{adv}} \quad \|\mathbf{x} - \mathbf{x}_t^{adv}\|_l \quad s.t. \quad h(\mathbf{x}_t^{adv}, t) \neq y_t$$

Notice that in contrast to eq. (1), def. 1 admits the existence of two adversarial patients, one for each treatment group (Fig. 2). Let us define the loss function on each group as the cross-entropy loss $J(\hat{f}(\mathbf{x}, t), y_t) = -y_t \log \hat{f}(\mathbf{x}, t) - (1 - y_t) \log(1 - \hat{f}(\mathbf{x}, t)), t \in \{0, 1\}$. Then, restating the optimisation problem as a maximisation of the loss results in the following modification of eq. (2):

$$\arg\max_{\mathbf{x}_t^{adv}} \quad J(\hat{f}(\mathbf{x}_t^{adv}, t), y_t) \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{x}_t^{adv}\|_\infty \leq \theta \quad \text{for} \quad t \in \{0, 1\} \tag{4}$$

Optimisation of eq. (4) can lead to a fast way of creating adversarial patients with respect to their potential outcomes by taking the first-order Taylor expansion of $J(\cdot)$ and keeping a fixed $\|\cdot\|_\infty$ perturbation. One advantage of phrasing the adversarial problem in terms of the loss $J(\cdot)$ is that it allows us to generalise to other functions, leading to a family of adversarial directions. Consequently, we can identify a general adversarial direction that jointly maximises the loss on both groups. To show that, let us define the joint loss as the convex combination of the cross-entropy losses on each group separately:

$$J_{joint}(\hat{f}(\mathbf{x},1), y_1, \hat{f}(\mathbf{x},0), y_0) = \pi \cdot J(\hat{f}(\mathbf{x},1), y_1) + (1-\pi) \cdot J(\hat{f}(\mathbf{x},0), y_0) \quad (5)$$

where $\pi \in [0,1]$ is a constant. We conjecture that π controls the relative importance of each loss and therefore can lead to adapting adversarial patients in a cost-sensitive framework (we will discuss this further in Sect. 6). An adversarial patient with respect to the joint loss will be derived by modifying an initial patient $\mathbf{x}$ to the direction that decreases the confidence of the model on making the correct prediction for both tasks (Fig. 3(a)). Indeed, differentiating eq. (5) with respect to $\mathbf{x}$ will result in the following expression:

$$\begin{aligned} &\nabla_{\mathbf{x}} J_{joint}(\cdot) = \pi \cdot \beta_{y_1} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x},1) + (1-\pi) \cdot \beta_{y_0} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x},0), \\ &\text{where} \quad \beta_{y_t=1} = -\frac{1}{\hat{f}(\mathbf{x},t)} \quad \text{and} \quad \beta_{y_t=0} = \frac{1}{1-\hat{f}(\mathbf{x},t)} \end{aligned} \quad (6)$$

Notice that $\beta_{y_t=1} < 0$ and $\beta_{y_t=0} > 0$, since $\hat{f}(\mathbf{x},t) \in (0,1)$. Suppose, for example, that $y_1 = 0$ and $y_0 = 1$. Then $\nabla_{\mathbf{x}} J_{joint}(\cdot) = \pi \cdot \beta_{y_1=0} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x},1) + (1-\pi) \cdot \beta_{y_0=1} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x},0)$ and the adversarial patient will be in a direction indicated by the weighted difference between the two gradients. In the next section, we show that stating the optimisation problem in terms of maximising a loss function results in identifying a second form of adversarial patients – those that can affect ITE.

4.2 Adversarial Patients and ITE

Let us continue with the hypothetical scenario of having knowledge of both potential outcomes. We will show that in counterfactual models there exist new types of adversarial directions, that are not a direct extension of existing definitions of adversarial examples. To see that let us define first the region, in which we will search for adversarial patients as $R(\mathbf{x},\theta) = \{\mathbf{x}' : \|\mathbf{x}-\mathbf{x}'\|_l \leq \theta\},\ \theta > 0$. For what follows we will consider $R(\cdot) = \|\cdot\|_\infty$, but other measures of distance can also be used. Suppose that for a patient $\mathbf{x}$ the potential outcomes are $y_1 = 1$ and $y_0 = 0$ and we have a model $\hat{f}$ that correctly assigns these outcomes with probabilities $\hat{f}(\mathbf{x},1) = 0.9$ and $\hat{f}(\mathbf{x},0) = 0.2$. Now consider a patient $\mathbf{x}'$ that maximises the joint loss on the potential outcomes within a region $R(\mathbf{x},\theta)$ and has the same potential outcomes as $\mathbf{x}$ but the model assigns them with probabilities $\hat{f}(\mathbf{x}',1) = 0.6$ and $\hat{f}(\mathbf{x}',0) = 0.4$. In this case $\mathbf{x}$ will have the same potential outcomes within the region $R(\mathbf{x},\theta)$ but in terms of the estimated $\hat{\text{ITE}}$ there is a large difference between $\mathbf{x}$ and $\mathbf{x}'$ ($\hat{\text{ITE}}(\mathbf{x}) = 0.7$ and $\hat{\text{ITE}}(\mathbf{x}') = 0.2$). In certain

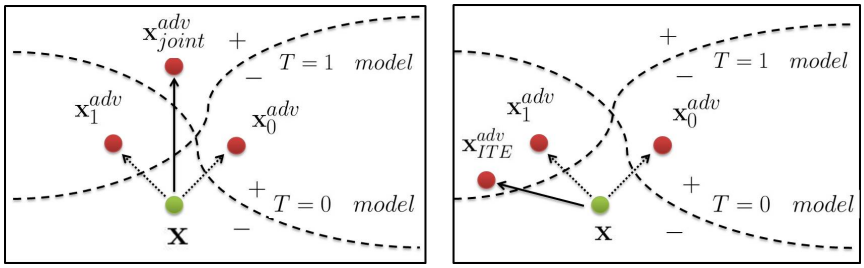


Fig. 3. Counterfactual models introduce, at least two, additional adversarial directions. An adversarial direction with respect to the joint loss on both groups will result in an adversarial patient $\mathbf{x}^{adv}_{joint}$ that would harm both models simultaneously (a). An adversarial direction with respect to ITE will result in an adversarial patient $\mathbf{x}^{adv}_{ITE}$ that maximally affects the difference between the potential outcomes (b).

tasks, such as subgroup identification, where the smoothness of ITE is crucial, $\mathbf{x}'$ can be considered as adversarial.

Maximisation of a loss on $\hat{\text{ITE}}$ within a fixed area of interest, $R(\mathbf{x}, \theta)$, will allow us to identify edge cases that may result to significantly different estimations. To quantify the loss on $\hat{\text{ITE}}$ we can adopt the ε_{PEHE}, which was defined in eq. (3). The empirical $\hat{\varepsilon}_{PEHE}$ on a patient $\mathbf{x}$ will be:

$$\hat{\varepsilon}_{PEHE}(\hat{f}(\mathbf{x},1), y_1, \hat{f}(\mathbf{x},0), y_0) = ((\hat{f}(\mathbf{x},1) - \hat{f}(\mathbf{x},0)) - (y_1 - y_0))^2 \tag{7}$$

If there is a treatment effect within the region $R(\mathbf{x}, \theta)$, i.e. the equality $\hat{f}(\mathbf{x}', 1) = \hat{f}(\mathbf{x}', 0), \forall \mathbf{x}' \in R(\mathbf{x}, \theta)$ does not hold, then there may be two edge cases: at least one example that maximally increases $\hat{\text{ITE}}(\mathbf{x})$ and/or at least one that maximally decreases it. Maximisation of eq. (7) with respect to $\mathbf{x}$ results in identifying the worst case perturbations with respect to the true ITE. We can define the new set of adversarial patients as follows.

Definition 2 (Adversarial Patient and ITE). *An adversarial patient with respect to ITE, $\mathbf{x}^{adv}_{ITE} \in \mathbb{R}^d$, is a solution to the following optimisation problem:*

$$\underset{\mathbf{x}^{adv}_{ITE}}{\arg\max} \quad \hat{\varepsilon}_{PEHE}(\hat{f}(\mathbf{x}^{adv}_{ITE}, 1), y_1, \hat{f}(\mathbf{x}^{adv}_{ITE}, 0), y_0) \quad s.t. \quad \mathbf{x}^{adv}_{ITE} \in R(\mathbf{x}, \theta)$$

Differentiating eq. (7) with respect to $\mathbf{x}$ will result in the following expression, where we have omitted the constants:

$$\begin{aligned} &\nabla_{\mathbf{x}} \hat{\varepsilon}_{PEHE}(\cdot) = \gamma_{y_1 - y_0} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x}, 1) - \gamma_{y_1 - y_0} \cdot \nabla_{\mathbf{x}} \hat{f}(\mathbf{x}, 0), \\ &\text{where} \quad \gamma_1 = \hat{f}(\mathbf{x},1) - \hat{f}(\mathbf{x},0) - 1, \quad \gamma_{-1} = \hat{f}(\mathbf{x},1) - \hat{f}(\mathbf{x},0) + 1 \\ &\text{and} \quad \gamma_0 = \hat{f}(\mathbf{x},1) - \hat{f}(\mathbf{x},0) \end{aligned} \tag{8}$$

Notice that since $\hat{f}(\mathbf{x}, t) \in (0, 1)$, we will have $\gamma_1 < 0$, $\gamma_{-1} > 0$ and $\gamma_0 = \hat{ITE}$. Comparison of eq. (6) and eq. (8) reveals that when $y_1 \neq y_0$ both directions lead

in decreasing the confidence of the two models on making the correct predictions. If $y_1 = y_0$ then the sign of the adversarial direction will depend on $\hat{\text{ITE}}$. For example if, for a patient $\mathbf{x}$, $\hat{\text{ITE}}(\mathbf{x}) > 0$, then an adversarial patient will be modified to a direction that increases it. In this case the model assigns the wrong treatment effect with higher confidence (Fig. 3(b)).

To summarise, in counterfactual models there exist, at least, the following adversarial patients, with potentially different consequences:

1. Adversarial patients with respect to the treatment model result in potentially different predictions about the benefit under the administered treatment (e.g. chemotherapy) irrespectively of the control model.
2. Adversarial patients with respect to the control model result in potentially different predictions about the benefit under the baseline treatment (e.g. standard care) irrespectively of the treatment model.
3. A family of adversarial patients that maximise the joint loss on both models result in potentially different predictions about the joint outcome (e.g. simultaneously changing the predictions about the effect of chemotherapy and standard care).
4. Adversarial patients with respect to ITE maximise the loss on the difference between the potential outcomes leading to wrong estimations of ITE with higher confidence, which may affect treatment decisions and lead to undesirable consequences (e.g. financial).

In the following section we verify the existence of the different adversarial directions and describe which of those can be observed or approximated as well as their implications in practical applications.

5 Adversarial Patients in Practice

So far we considered the hypothetical scenario of observing both potential outcomes. This allowed us to identify the existence of new adversarial directions that challenge the local stability of the potential outcomes and the local smoothness of ITE. We now turn to the realistic case, where we only observe one of the outcomes. In practice, among the two directions that can deceive each model (Fig. 2), we can reliably identify only the one that corresponds to the observed outcome. Algorithm 1 describes how to construct adversarial patients with iterative FGSM [9] for a model $\hat{f}(\cdot, t)$ using projected gradient ascent.

In order to perform adversarial training using adversarial patients that require the counterfactual outcome, we will adopt a nearest neighbour model by taking into account relevant factual outcomes. Nearest neighbour approximations of $\hat{\varepsilon}_{PEHE}$ have also been adopted in recent works, either directly as part of the objective function [8], or as a measure of the generalisation performance for hyperparameter selection [17]. We will focus on two types of adversarial patients – those that affect the loss on the observed (factual) outcome and those that affect ITE, using the nearest neighbour approximation.

Algorithm 1 Generating Adversarial Patients

Input: Patient $\mathbf{x}$, factual outcome y_t, model $\hat{f}(\cdot, t)$, perturbation θ, step α, iterations m
Output: Adversarial patient $\mathbf{x}_t^{adv}$

1: $\mathbf{x}_t^{adv} = \mathbf{x}$
2: **for** $i := 1$ to m **do**
3: $\quad \mathbf{x}_t^{adv} = \mathbf{x}_t^{adv} + \alpha \cdot sign(\nabla_{\mathbf{x}} J(\hat{f}(\mathbf{x}_t^{adv}, t), y_t))$
4: $\quad$ **if** $\|\mathbf{x}_t^{adv} - \mathbf{x}\|_\infty > \theta$ **then**
5: $\quad\quad$ Project $\mathbf{x}_t^{adv}$ onto the boundary of the feasible region
6: **end for**
7: Return $\mathbf{x}_t^{adv}$

To verify the effect of the different types of adversarial directions we consider the following outcome function [5], which we will call Simulated Model 1 (SM1):

$$\text{SM1} : \text{logit}(f(\mathbf{X}, T)) = -1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.5X_2X_7 + \lambda T \mathbb{1}(\mathbf{x} \in S)$$

where the features X_j are drawn independently from a normal distribution, $X_j \sim N(0, 1), j = 1, ..., 15$. We choose $\lambda = 2$ and S as the region $X_0 < 0.545 \cap X_1 > -0.545$. We create a sample of 2000 examples and we average the results over 100 realisations of the outcome with 50/25/25 train/validation/test splits. To create adversarial patients we use alg. 1 where we also considered modifications of line 3 by substituting the factual loss with: 1. $\hat{\varepsilon}_{PEHE}$ and 2. a nearest neighbour approximation of $\hat{\varepsilon}_{PEHE}$. Figure 4 verifies the different effect of each adversarial direction. Notice that adversarial patients that maximise $\hat{\varepsilon}_{PEHE}$ can have an impact on the factual loss (and vice-versa). However their effect will be small since these directions do not always cause the examples to cross the decision boundary. In fact they can even make the model to be more confident for a correctly predicted factual outcome.

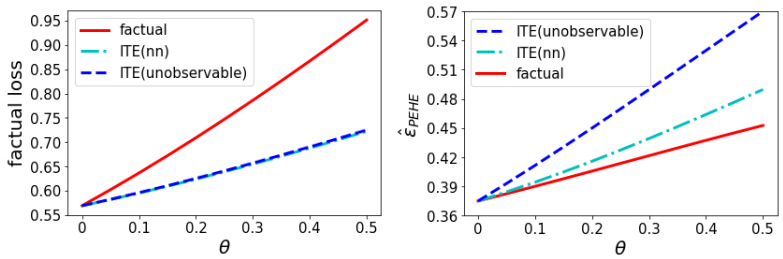


Fig. 4. In counterfactual models we can identify adversarial patients with respect to the potential outcomes and with respect to ITE. To show the existence of these different types we create adversarial patients by maximising the factual loss (factual), $\hat{\varepsilon}_{PEHE}$ (ITE) and a nearest neighbour approximation of $\hat{\varepsilon}_{PEHE}$ (ITE(nn)). We report the effect of the different types of adversarial patients on the two loss functions as the size of the search area $R(\mathbf{x}, \theta)$ increases.

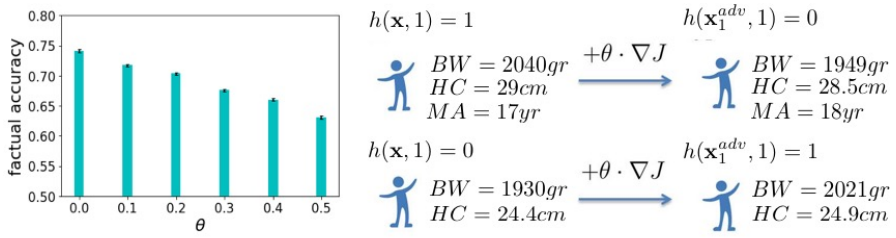


Fig. 5. Adversarial patients can deceive deep counterfactual models. Here we observe that small, worst-case perturbations (in only 2 or 3 of the 25 features) result in a different predicted outcome, that may be in contrast to clinical intuition.

5.1 Case Study: Adversarial Patients as Warning Flags

To show the effect of adversarial patients with respect to their factual outcome in a real-world scenario we consider the Infant Health and Development Program (IHDP) [7]. The dataset is from a randomised experiment evaluating the effect of specialist home visits on children's cognitive test scores. It comprises of 985 subjects and 25 features (6 continuous, 19 binary) measuring characteristics of children and their mothers. For the outcome function we used the response surface B of Hill [7] to form a binary classification task. We averaged the results over 1000 realisations of the outcome with 65/25/10 train/validation/test split and report the test set accuracy on the observed (factual) outcome (Fig. 5). We used a network with 2 shared layers and 2 group-specific with 50 nodes each [1].

To ensure *interpretable* adversarial patients, we restrict ourselves to an area $R(\mathbf{x}, \theta)$ modifying only three features: birth weight (BW), head circumference (HC) and mother's age (MA), and round each to the closest value it can take according to its domain. We observe in Fig. 5 that a small modification (we modify only 3 features, and fix the maximum perturbation to a minimal value) is enough to create adversarial patients that deceive the model. Consider the second patient in Fig. 5. We know that if the child receives the treatment ($t = 1$) she will not have a positive outcome ($y_1 = 0$). However, for a child $\mathbf{x}_1^{adv}$ with two slightly different features the model would predict the outcome will be positive ($h(\mathbf{x}_1^{adv}, 1) = 1$). Such small perturbations being responsible for potentially life-changing decisions is a significant ethical dilemma. Adversarial patients may result in different treatment decisions, a taxonomy of which is given in table 1. In our example, for the initial patient the treatment could be either ineffective or harmful depending on the counterfactual outcome. A small perturbation results in concluding that the treatment is either essential or unnecessary. Such a small perturbation may lead to a different treatment decision, such as "treat" or "no further action required", while in reality the right decision could be "do not treat". A detailed study of adversarial patients and their consequences in the context of treatment decisions is imperative and will be the focus of future work.

[1] Further details on the outcome function and evaluation protocol can be found in the supplementary material.

Table 1. Each pair of potential outcomes defines an action that needs to be taken. Suppose that we are in scenario A and our model correctly predicts that the patient needs to get the treatment. An adversarial patient could be the minimal perturbation required to arrive to a different treatment decision. For example, the minimum perturbation required to force the decision "Do not treat" would be the one that changes the predicted outcome of both models (scenario D).

Scenario	Treatment Status	Prognosis	Possible Decision
A. $y_1 = 1, y_0 = 0$	Essential	Negative	Treat
B. $y_1 = 1, y_0 = 1$	Unnecessary	Positive	No further action
C. $y_1 = 0, y_0 = 0$	Ineffective	Negative	Search for alternative
D. $y_1 = 0, y_0 = 1$	Harmful	Positive	Do not treat

5.2 Case Study: Subgroup Identification

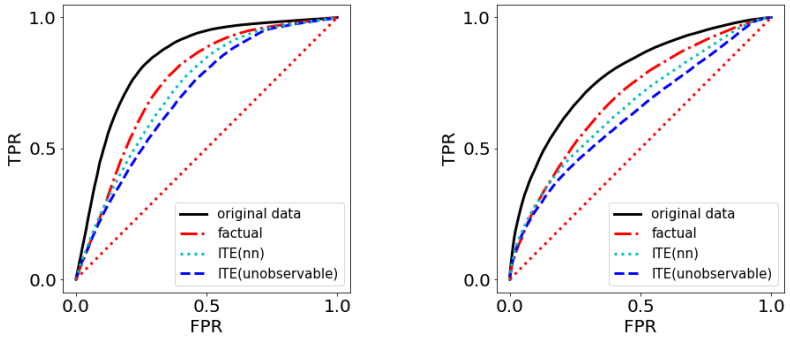
Exploratory subgroup identification is a critical task in phase III trials. Subgroups are usually defined in terms of a few features[1] associated with a trial population with some desirable property, such as improved treatment effect. As described by Sechidis et al. [16], the outcome can be realised as a function of two types of features, each one providing useful information for different tasks:

1. *prognostic* features that affect the outcome irrespectively of the treatment.
2. *predictive* features that affect the outcome through their interaction with the treatment.

Prognostic features are important for clinical trial planning as they can be considered for stratification and randomisation [14]. Predictive features are important for treatment effect estimation. A significant body of literature has focused on finding subgroups of patients with enhanced treatment effect [10]. Formally a subgroup $\hat{S}$ of enhanced treatment effect includes all patients for which $\hat{\text{ITE}}(\mathbf{x}) > \delta,\ \forall \mathbf{x} \in \hat{S}$, where δ is a constant chosen based on the clinical setting.

To show how multi-task counterfactual NNs can be used for subgroup identification we again consider our simulated model SM1. In this model X_1 and X_2 have both prognostic and predictive effect, while X_7 is solely prognostic – i.e. it affects the outcome but not ITE. We also consider a modification of SM1 where the subgroup is defined as the region $X_3 < 0.545 \cap X_4 > -0.545$ (we will refer to this case as SM2). In this case X_3, X_4 are solely predictive, while X_1, X_2, X_7 will have only a prognostic effect. In this case, ITE should be influenced only by X_3 and X_4. We evaluate the performance on subgroup identification using true positive/false positive rates while varying the threshold δ (a patient $\mathbf{x}$ is included in the subgroup if $\hat{\text{ITE}}(\mathbf{x}) > \delta$) and comparing the resulting subgroup, $\hat{S}$, with the ground truth S. The datasets were created similarly to sect. 5 and the results were averaged over 100 realisations of the outcome.

[1] Since a subgroup needs to be interpretable for clinicians, they are defined in terms of the features with the strongest predictive effect (usually less than 3 features) [5].



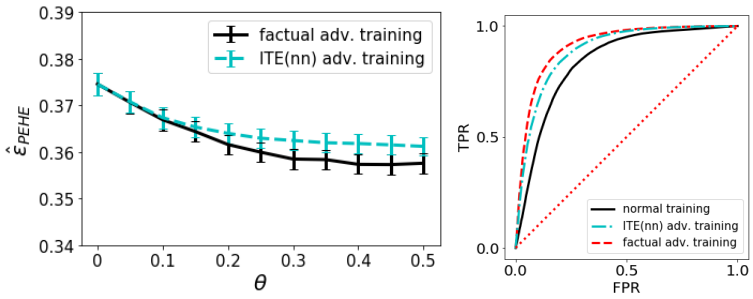
(a) Adversarial examples for SM1 (b) Adversarial examples for SM2

Fig. 6. Adversarial patients with identical predictive features can have significantly different estimated treatment effects. We created adversarial patients ($\theta = 0.4$) with respect to the factual outcome (factual) and with respect to ITE using the nearest neighbour approximation (ITE(nn)). We also report results for examples created with the true counterfactual outcome (ITE) for comparison. Different estimations of ITE may result in incorrectly removing(adding) patients from(to) the subgroup.

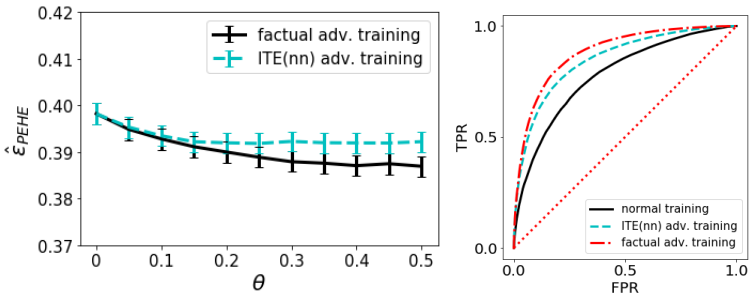
To see whether the estimated $\hat{\text{ITE}}$ can be influenced by non-predictive features, we created adversarial patients that are identical with respect to their predictive features. Figure 6 shows that adversarial patients can deceive the model to wrongly remove(add) them from(to) the subgroup, since the TPR/FPR curves are closer to the centre diagonal line than the original. Notice that adversarial patients with respect to ITE have higher effect (i.e. closer to the diagonal) than adversarial patients with respect to the factual outcome. Our results validate that patients with identical predictive features can have different $\hat{\text{ITE}}$.

We can counter this bias with an adversarial training approach, making the model robust to change in non-predictive features. We create adversarial patients with respect to their factual outcome and with respect to ITE using the nearest neighbour approximation of $\hat{\varepsilon}_{PEHE}$. We restrict the area of search $R(\mathbf{x}, \theta)$ by keeping constant the two features with the highest predictive strength. Following Goodfellow et al. [6] we use equal weight on legitimate/adversarial examples.

We observe that adversarial training acts as a regulariser improving the generalisation performance, measured by $\hat{\varepsilon}_{PEHE}$ (left side of Fig. 7). On the right of Fig. 7 we see that adversarial training can also improve the performance on the task of subgroup identification. We also observe that adversarial training with respect to the factual outcome acts as a stronger regulariser. Therefore training the model to make similar predictions for patients differing only on their prognostic/irrelevant part leads to better estimates of ITE. Note that here we assumed knowledge of the predictive features, which can be based on suggestions from domain experts or derived algorithmically [16]. In practice it is also common to know at least a subset of the prognostic features, as they are often used for stratification (among other tasks) [14]. In the supplementary material we present additional results: e.g. modify all features, and more simulated models.



(a) Effect of adversarial training on SM1



(b) Effect of adversarial training on SM2

Fig. 7. Adversarial training improves the generalisation of counterfactual models. We trained a simple counterfactual NN on two outcome functions and report the estimated error $\hat{\epsilon}_{PEHE}$ (left column) and TPR/FPR curves for $\theta = 0.25$ (right column). We observe that adversarial training with respect to the factual outcome acts as a strong regulariser for ITE estimation. The effect of adversarial training is attributed to the model becoming invariant with respect to changes on prognostic or irrelevant features.

6 Discussion & Future Research

The study of adversarial examples in the context of clinical trials, raises several challenges for research. Such small perturbations being responsible for potentially life-changing decisions is a significant ethical issue. Here we highlight a "wish-list" of issues we believe could be addressed by the community.

New problems

- How can we ensure medically plausible and clinically interpretable adversarial patients?
- How can we reliably create adversarial patients with respect to population level metrics such as average treatment effect, which may have influence on issues like drug pricing?
- To the best of our knowledge, this paper is the first to tackle subgroup identification using NNs. An in-depth study of adversarial patients for robust subgroup identification is likely to be a fruitful area of work.

- Medical regulatory bodies demand strong evidence, ideally guarantees, not conjectures. What properties can we prove about adversarial patients, e.g. generalisation bounds? How should policy makers regard this?

Ethical issues

- Healthcare is rife with cost-sensitive decision making—with imbalanced ethical, personal and financial costs, for most choices that arise. How can we identify/use adversarial patients with this in mind?
- Identifying adversarial directions common to several classes of model may be suitable to influence policy—for example if we find particular health characteristics/features to be sensitive, they should be treated with care when recording from patients.

Making use of, and contributing to medical knowledge

- How can we build domain-specific medical knowledge (e.g. co-morbidities or known metabolic pathways where the drug is targeted) into the identification of adversarial patients to be used as warning flags?
- Our work has highlighted that we can have prognostic or predictive adversarial directions. How can we use this knowledge to better identity trustworthy predictive biomarkers for reliable treatment assignments?
- Outcomes of trials are often *structured*, e.g. multiple correlated and/or hierarchical outcomes. How can we adapt adversarial methods for this?

Technical issues

- A defining aspect of adversarial patients is they can be created with respect to an approximation of the ground truth label (e.g. ITE with nearest neighbours). What is the influence of this on the quality/plausibility of the created adversarial examples?
- The adversarial methodologies we use (e.g. FGSM, adversarial training) are for purely illustrative purposes. There is a significant body of literature (e.g. [12, 11, 19]) that could be adopted in this context.

7 Conclusions

We have studied the idea of adversarial examples in counterfactual models, for personalised medicine. The concept of "adversarial patients", can exist in many forms—we have shown that some adversarial directions cannot be observed, but still provably exist. We showed that small input perturbations (but still medically plausible) of multi-task counterfactual networks can lead to predictions that may not be in accordance with human intuition. Training a model to account for them can affect critical tasks in the clinical trial setting, such as exploratory subgroup identification. We propose that the study of adversarial patients in personalised medicine, where mispredictions can result in potentially life changing decisions, is an imperative research direction for the community to address.

Acknowledgments. K.P. was supported by the EPSRC through the Centre for Doctoral Training Grant [EP/1038099/1]. K.S. was funded by the AstraZeneca Data Science Fellowship at the University of Manchester. G.B. was supported by the EPSRC LAMBDA project [EP/N035127/1].

References

1. Alaa, A.M., Weisz, M., van der Schaar, M.: Deep counterfactual networks with propensity-dropout. In: ICML Workshop on Principled Approaches to Deep learning (2017)
2. Athey, S., Imbens, G.: Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences **113**(27), 7353–7360 (2016)
3. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: ECML/PKDD. pp. 387–402. Springer (2013)
4. European Parliament and Council of the European Union: Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (data protection directive). L119 pp. 1–88 (2016)
5. Foster, J.C., Taylor, J.M., Ruberg, S.J.: Subgroup identification from randomized clinical trial data. Statistics in medicine **30**(24), 2867–2880 (2011)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
7. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011)
8. Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: ICML. pp. 3020–3029 (2016)
9. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: ICLR (2017)
10. Lipkovich, I., Dmitrienko, A., et al.: Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Statistics in medicine **36**(1), 136–196 (2017)
11. Miyato, T., Maeda, S.i., Koyama, M., Nakae, K., Ishii, S.: Distributional smoothing with virtual adversarial training. In: ICLR (2016)
12. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2574–2582 (2016)
13. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. pp. 372–387. IEEE (2016)
14. Ruberg, S.J., Shen, L.: Personalized medicine: Four perspectives of tailored medicine. Statistics in Biopharmaceutical Research **7**(3), 214–229 (2015)
15. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66**(5), 688 (1974)
16. Sechidis, K., Papangelou, K., Metcalfe, P.D., Svensson, D., Weatherall, J., Brown, G.: Distinguishing prognostic and predictive biomarkers: An information theoretic approach. Bioinformatics **1**, 12 (2018)
17. Shalit, U., Johansson, F., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: ICML (2017)
18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
19. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: ICLR (2018)
20. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association (just-accepted) (2017)