Journal of Biomedical Informatics 109 (2020) 103515

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The Counterfactual χ-GAN: Finding comparable cohorts in observational health data

Amelia J. Averitt[a,*], Natnicha Vanitchanant[a], Rajesh Ranganath[b], Adler J. Perotte[a]

[a] Biomedical Informatics, Columbia University, New York, NY, United States
[b] Courant Institute, Center for Data Science, New York University, New York, NY, United States

ARTICLE INFO

Keywords:
Causal inference
Machine learning
Health
GANs
Deep learning
Observational studies

ABSTRACT

Causal inference often relies on the counterfactual framework, which requires that treatment assignment is independent of the outcome, known as strong ignorability. Approaches to enforcing strong ignorability in causal analyses of observational data include weighting and matching methods. Effect estimates, such as the *average treatment effect* (ATE), are then estimated as expectations under the re-weighted or matched distribution, P. The choice of P is important and can impact the interpretation of the effect estimate and the variance of effect estimates. In this work, instead of specifying P, we learn a distribution that simultaneously maximizes coverage and minimizes variance of ATE estimates. In order to learn this distribution, this research proposes a generative adversarial network (GAN)-based model called the Counterfactual χ-GAN (cGAN), which also learns feature-balancing weights and supports unbiased causal estimation in the absence of unobserved confounding. Our model minimizes the Pearson χ^2-divergence, which we show simultaneously maximizes coverage and minimizes the variance of importance sampling estimates. To our knowledge, this is the first such application of the Pearson χ^2-divergence. We demonstrate the effectiveness of cGAN in achieving feature balance relative to established weighting methods in simulation and with real-world medical data.

1. Introduction

Counterfactual Causal Inference. In biomedicine, causal assessment often relies on the framework of *counterfactual inference*. This framework requires that causal effects are estimated by contrasting the distribution of outcomes under different treatments (T) [9]. Under the counterfactual theory, each individual, i, has a *potential outcome* (Y_T^i) given that they received a treatment ($T = 1$) and a control ($T = 0$). For example, the treatment effect of metformin on the outcome of change in fasting blood glucose levels would be evaluated as the difference in values for the same individual when taking and not taking metformin. These potential outcomes are given by Y_1^i and Y_0^i, respectively. This framework seeks to contrast the outcome, Y for an individual under these two states [38]. The causal effect of the treatment on the outcome is then summarized by calculating population-level effect estimates, such as the average treatment effect (ATE). This is defined as the expected difference in outcomes over all individuals (Eq. 1).

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \tag{1}$$

However, the estimation of causal effects from these potential outcomes requires access to the outcome for the state in which units were not assigned. These are known as *counterfactuals*, as they are contrary to reality. In practice, counterfactuals are never observed as a single individual (or population) cannot simultaneously be both treated and untreated. This is known as the 'fundamental problem of causal inference.' As such, causal assessments may rely on approximations that employ an additional population that serves as a proxy for the unobserved states [16]. These approximations seek to construct populations such that the observed ATE, $\widehat{ATE}$, equals the true ATE that would arise from a counterfactual population. In other words, we seek an $\widehat{ATE}$ that is *unbiased*.

A sufficient condition for unbiased $\widehat{ATE}$ estimation is that $\mathbb{E}[Y_1|T = 1] = E[Y_1|T = 0]$ and $\mathbb{E}[Y_0|T = 0] = E[Y_0|T = 1]$ [25]. Within the counterfactual framework, this equality is central to the assumption of *strong ignorability* (Eq. 2) [35].

$$Y_i(1),\ Y_i(0) \perp\!\!\!\perp T_i \tag{2}$$

This assumption states a unit's assignment to a treatment is independent of that unit's potential outcomes, Y_i, and that treatment assignment is, therefore, ignorable. Causal claims borne from data that satisfy this requirement are regarded as unbiased as all confounding factors that could induce a dependence between Y_i and T_i are equally

* Corresponding author.
E-mail address: amelia.averitt@gmail.com (A.J. Averitt).

https://doi.org/10.1016/j.jbi.2020.103515
Received 9 January 2020; Received in revised form 15 July 2020; Accepted 16 July 2020
Available online 07 August 2020

represented in the treatment and comparator arms [38]. Consequently, this means that the distribution of features is the same in both arms and features are said to be *balanced*. Other assumptions, such as positivity and the Stable Unit Treatment Value Assumption (SUTVA), are also necessary and assumed to be true [39].

Counterfactual Inference from Biomedical Data. In biomedicine, the current gold-standard for upholding strong ignorability in counterfactual inference is randomized experimentation, wherein the random allocation of study units to treatment or control arms eliminates confounding.[38]. However, randomized experiments, including randomized controlled trials (RCTs), may be expensive, time-consuming, and ethically fraught [6]. Furthermore, RCTs often enforce unrealistic assumptions that impede the generalization of causal knowledge to the real world [4]. Observational data is regarded as a more externally-valid source from which to generate causal knowledge but, in the absence of randomization to treatment, strong ignorability cannot be guaranteed. As such, causal claims from observational data sources may be spurious or misleading. When not randomized, observational data may be manipulated such that the strong ignorability assumption is upheld and causal claims are unbiased in the absence of unobserved confounders.

Matching and weighting are popular pre-analysis manipulations to approximate the unconditional form of strong ignorability in observational populations. These methods create pseudo-populations in which the assumption is met without need for further manipulation [37]. This is opposed to methods of statistical adjustment, which occur peri-analysis, and approximate the conditional form of strong ignorability [27]. Arguably, the most common strategy for weighting is the *inverse probability of treatment weighting* (IPW) [43], though other methods include the direct minimization of imbalance [10,19,20] or weighting by the odds of treatment, kernel weighting, and overlap weighting [36,15,14,28,22].

Limitations to Counterfactual Inference from Observational Data. A commonality among these methods for observational data is that they implicitly or explicitly all specify a distribution function, P, that the expectation in Eq. 1 is taken with respect to. This distribution is often the distribution associated with the treated ($p_1(x)$), the controls ($p_2(x)$), or a combination thereof (e.g. $\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)$). However, this choice of distribution can lead to high variance effect estimates in circumstances where there are regions of poor overlapping support between the treated and untreated populations. An effect of this is often observed in the context of IPW analyses with propensity scores near zero or one [23]. This may result in unstable downstream causal effect estimates and poor coverage of feature space, which may impact the validity of the causal effect estimate.

Finding Comparable Cohorts in Observational Data. In this work, we instead construct an implicit distribution, P, that focuses on the regions of the sample space with significant overlap between the treated and untreated populations. Such a construction involves an inherent trade-off between coverage and variance. We propose the Counterfactual χ-GAN (cGAN) that uses an adversarial approach to learn stable, feature-balancing weights without reliance on the propensity score. The cGAN utilizes a unique architecture which identifies a target distribution that minimizes the importance sampling variance for approximations of $\mathbb{E}[Y_1|T=1]$ and $\mathbb{E}[Y_0|T=0]$. This objective simultaneously encourages coverage and identifies the importance sampling weights that result in pseudo-populations that satisfy the unconditional form of strong ignorability.

This paper proceeds as follows: Section 2 defines the model and learning procedures, Section 3 presents an evaluation of this model through a simulation and an application to real-world clinical data, and finally, Section 4 present the results, and Section 5 discusses open issues, limitations of the model, and future work.

2. The model

We introduce the *Counterfactual χ-GAN* (cGAN), an adversarial approach to feature balance in causal inference that is based on importance sampling theory. Using an adversarial approach based on variational minimization based on the *f*-GAN, we minimize the sum of the Pearson χ^2-divergences between a deep generative model and the sampling distributions from each arm of a study. We show that minimizing the χ^2-divergence is equivalent, up to a constant factor, to minimizing the variance of importance sampling estimates to be made in approximating quantities such as ATEs. Similar to other weighting approaches, this approach assumes SUTVA, positivity, and no unmeasured confounders. In the following, P is the constructed target distribution and Q_a is the sampling distribution for each study arm.

2.1. Importance sampling and the χ^2-divergence

Importance sampling is a strategy for estimating expectations under an unknown target distribution given a known proposal distribution [31]. Though the importance sampling has broader usage than our application, we focused on the use of importance sampling for estimation of the average treatment effect (ATE) because of its close relationship with the χ^2 divergence. The importance sampling weight is defined as a likelihood ratio: the likelihood of an observation under the target distribution, $p(x)$ divided by the likelihood under the proposal distribution, $q(x)$. Weighted expectations based on the proposal distribution approximate unweighted expectations from the target distribution at shown in Eq. 3.

$$\mathbb{E}_q\left[\frac{p(x)}{q(x)}\phi(x)\right] = \mathbb{E}_p[\phi(x)] \tag{3}$$

Consider the units in an arm of an observational study as being samples from such a proposal distribution. One strategy for obtaining unbiased expectations of treatment effects is to identify importance sampling weights for each arm that approximate expectations from a shared target distribution. However, this problem is underspecified given that we could choose any target distribution with the correct support. In this work, we choose the target distribution that yields importance sampling approximations with smallest variance. Eq. 4 shows the form for the variance of importance sampling estimates where $\phi(x)$ is the constant function. This choice is to make the formulation of the cGAN as outcome agnostic as possible. This form highlights its connection with the χ^2-divergence, which has a function form as shown in Eq. 5. This connection was previously noted in [5]. Therefore, the solution which minimizes the χ^2-divergence would also minimize the variance of expectations for unknown outcomes. Of note, importance sampling is known to be a method that can produce high variance estimates, but since we will be minimizing the variance directly, this is less of a concern here.

$$\sigma_q^2 = \frac{\mu^2}{n}\left(\int q(x)\left[\frac{p(x)^2}{q(x)^2} - 1\right]\right)dx \tag{4}$$

$$\chi^2 p||q = \int q(x)\left[\frac{p(x)^2}{q(x)^2} - 1\right]dx \tag{5}$$

2.2. Likelihood ratios, overlap, & the ATE

Importance sampling weights can be leveraged to estimate an ATE in that region of $q(x)$ where there is significant overlap of probability mass/density between treatment arms. This is the region that satisfies the idea of a natural experiment and in which ATE estimations are reliable. Informally, we seek to get the most coverage of the overlapping region of $q(x)$, as it results in importance sampling estimates with low variance.

Typically, the expectation in the ATE is taken with respect to the original feature distribution, $q(x)$. Under cGAN-weighted data, expectations are taken with respect to the target distribution $p(x)$. As such, calculations of the ATE from the cGAN are not equivalent to what

many would classically consider the ATE, but rather, is an ATE with respect to the new, learned feature distribution. We call this new estimate the ATE_p. This inequality is demonstrated in Eq. 6. This set of equations shows that the typical ATE, ATE_q, is not equivalent to the expectation that we estimate, the ATE_p.

$$\begin{aligned} ATE_q &= \mathbb{E}_{q(y_1)}[y_1] - \mathbb{E}_{q(y_0)}[y_0] \\ &= \mathbb{E}_{q(x)}\mathbb{E}_{q(y_1|x)}[y_1|x] - \mathbb{E}_{q(x)}\mathbb{E}_{q(y_0|x)}[y_0|x] \\ &= \mathbb{E}_{q(x)}\mathbb{E}_{q(y|x,t=1)}[y|x,\ t=1] - \mathbb{E}_{q(x)}\mathbb{E}_{q(y|x,t=0)}[y|x,\ t=0] \\ &= \mathbb{E}_{q(x|t=1)}\frac{q(x)}{q(x\mid t=1)}\mathbb{E}_{q(y|x,t=1)}[y|x,\ t=1] \\ &\quad - \mathbb{E}_{q(x|t=0)}\frac{q(x)}{q(x\mid t=0)}\mathbb{E}_{q(y|x,t=0)}[y|x,\ t=0] \\ &\neq \mathbb{E}_{q(x|t=1)}\frac{\mathbf{p}(x)}{q(x\mid t=1)}\mathbb{E}_{q(y|x,t=1)}[y|x,\ t=1] \\ &\quad - \mathbb{E}_{q(x|t=0)}\frac{\mathbf{p}(x)}{q(x\mid t=0)}\mathbb{E}_{q(y|x,t=0)}[y|x,\ t=0] \end{aligned} \tag{6}$$

Consider two distributions Q_1 and Q_2 that represent two arms of a study. It is possible to make unbiased ATE_p estimates based on a single distribution, P, leveraging likelihood ratios/importance sampling weights as shown in Eq. 7.

$$ATE_p = \mathbb{E}_p[Y_1] - \mathbb{E}_p[Y_0] = \mathbb{E}_{q_1}\left[\frac{p(x)}{q_1(x)}Y_1\right] - \mathbb{E}_{q_2}\left[\frac{p(x)}{q_2(x)}Y_0\right] \tag{7}$$

We will leverage an approach based on adversarial learning to simultaneously maximizes coverage, minimizes the variance defined in Eq. 4, and directly estimates likelihood ratios, $\frac{p(x)}{q_1(x)}$ and $\frac{p(x)}{q_2(x)}$.

2.3. f-GAN

The f-GAN framework provides a strategy for estimation and minimization of arbitrary f-divergences based on a variational divergence minimization approach [32].

$$D_f P\|Q = \int_{\mathcal{X}} q(x) \sup_{t\in dom_{f^*}} \left\{t\frac{p(x)}{q(x)} - f^*(t)\right\}dx \tag{8}$$

$$\geqslant \sup_{T\in\mathcal{T}} \left(\int_{\mathcal{X}} p(x)T(x)dx - \int_{\mathcal{X}} q(x)f^*(T(x))dx\right) \tag{9}$$

$$= \sup_{T\in\mathcal{T}} \left(\mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))]\right) \tag{10}$$

where T is a class of function such that T: $\mathcal{X} \to \mathbb{R}$, f is the function that characterizes the χ^2-divergence, $f(u) = (u-1)^2$, f^* is the Fenchel conjugate of f, $f^*(t) = \frac{1}{4}t^2 + t$, and P and Q are probability distributions with continuous densities, $p(x)$ and $q(x)$. T is typically a multilayer neural network. This formulation lower bounds the χ^2-divergence based on functions T, P, and Q in such a way that unbiased noisy gradients of the lower bound can be easily obtained based on samples from P and Q. In addition, the variational function, T, has a tight bound for $T^* = f'\left(\frac{p(x)}{q(x)}\right)$ which is equivalent to $2\left(\frac{p(x)}{q(x)} - 1\right)$ in the case of the χ^2-divergence. To respect the bounds of T that result in valid likelihood ratios, we represent T as a nonlinear transformation of an unbounded function V: $T(x) = g_f(V(x)) = -2 + log(1 + e^{V(x)})$. The likelihood ratio, $\frac{p}{q}$, is easily derived from here and provides the importance sampling weights necessary for approximating expectations under $p(x)$ as shown in Eq. 3.

2.4. The counterfactual χ-GAN

The cGAN builds on importance sampling theory and extends the f-GAN framework to learn feature balancing weights through an adversarial training process. Previously, [42] have explored importance weights from critics of divergence-based GAN models. However, unlike this method and other f-GANs where there is a generator, G and a single variational function, the cGAN employs dual training from at least two variational functions (Fig. 1).

Consider a set of A treatments, each associated with one of A populations, or arms of a study. Each population contains N_a units and are drawn from an unknown and population-specific distribution Q_a. Based on the connection between the χ^2-divergence and the variance of importance sampling estimates outlined above, our objective is to identify a target distribution that minimizes the χ^2-divergence to all populations being compared: $\text{argmin}_p \sum_{a=1}^{A} \chi^2(p(x)\|q_a(x))$. This is the sum of the divergences between the generator and the unweighted treatment arms. It is minimized when $p(x)$ equals $q_a(x)$ for all a and is directly proportional to the sum of the variances of importance sampling estimates under the target distribution, P, with proposals, Q_a. Because of the constant in Eq. 4, minimizing the χ^2-divergence is equivalent to minimizing a normalized variance which weighs each population equally regardless of the number of units and the magnitude of the treatment effect, ϕ.

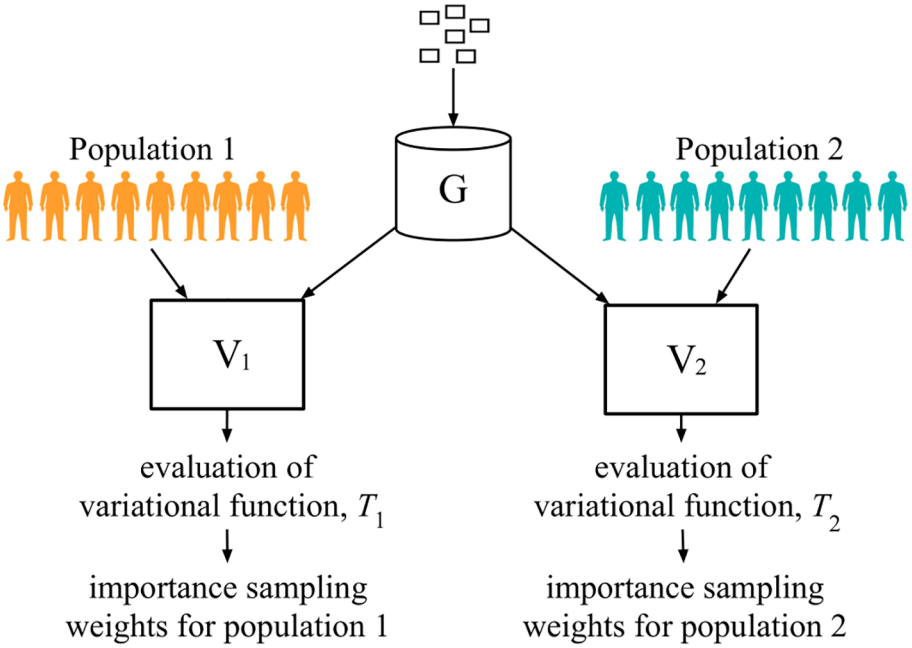

Fig. 1. Architecture of Counterfactual χ-GAN.

Algorithm 1. Minibatch stochastic gradient cGAN optimization

Input : $(x_{1,1},...,x_{1,N_1},...,x_{A,N_A})$

Output : $\theta, \omega_{1:A}$

Initialize θ, $\omega_{1:A}$ and minibatch size, M.

while $F(\theta, \omega_{1:A})$ *not converged* **do**

  for $a \in (1, \ldots, A)$ *treatment groups* **do**

    Sample a batch of noise samples, $z_{1:M} \sim p_g$, where p_g is a prior distribution such as an isotropic Gaussian

    Sample minibatch of data, $x_{a,1:M} \sim q_a$

    Compute gradient w.r.t. variational function parameters

$$\nabla_{\omega_a} F = \sum_{m=1}^{M} \nabla_{\omega_a} (g_f(V_{\omega_a}(G_\theta(z_m))) - \tfrac{1}{4} g_f(V_{\omega_a}(x_{a,m}))^2 - g_f(V_{\omega_a}(x_{a,m})))$$

    Ascend the ω_a gradient according to a gradient-based optimizer

  end

  Compute gradient w.r.t. generator parameters

$$\nabla_\theta F = \sum_{m=1}^{M} \sum_{a=1}^{A} \nabla_\theta \left[g_f(V_{\omega_a}(G_\theta(z_m))) \right]$$

  Descend the θ gradient according to a gradient-based optimizer

  Update V_{ω_a} and G_θ learning rates according to schedule

end

As a byproduct of minimizing this divergence, we will also identify a set of *importance weights*, $w_{a,n}$, for each unit in each population that allows estimation of expectations from the same target distribution, P, thus satisfying the unconditional form of strong ignorability. Using these importance weights, expectations can be approximated as $\mathbb{E}_p[f] \approx \sum_{n=1}^{N_a} w_a \phi(x_{a,n})$ where $w_{a,n} = \frac{1}{c} \frac{p(x_{a,n})}{q_a(x_{a,n})}$, where $c = \sum_{n=1}^{N_a} \frac{p(x_n)}{q_a(x_n)}$ is an normalizing constant, p is the density of the shared target distribution, q_a is the density of the proposal distribution, and $x_{a,n} \sim Q_a$. Note that our strategy eliminates the need to explicitly evaluate $p(x_{a,n})$ and $q_a(x_{a,n})$ as the likelihood ratio is estimated directly by the f-GAN. If desired, expectations can also be approximated using the sample-importance-resampling (SIR) algorithm where samples approximately distributed according to p can be simulated by drawing samples from the weighted empirical distribution $\hat{q}_a(x) = \frac{1}{N_a} \sum_{n=1}^{N_a} w_{a,n} \delta(x - x_{a,n})$ [7].

The objective function for the cGAN is shown in Eq. 11 and is closely related to the objective defined in [32]. θ parameterizes the generative model and ω_a parameterizes the variational model for each treatment arm, a. In our experiments, V_{ω_a} for all a are neural networks that mirror discriminators in the traditional GAN framework and P_θ is a neural networks that mirrors the generator. Note that the generator in the original f-GAN framework is usually Q_a. In our case, to achieve the desired directionality of the χ^2-divergence, the empirical distribution must be Q_a and the generator must be P.

$$F(\theta, \omega_{1:A}) = \sum_{t=1}^{A} \left(\mathbb{E}_{x \sim P_\theta}[g_f(V_{\omega_t}(x))] + \mathbb{E}_{x \sim Q_a}\left[-\frac{1}{4} g_f\left(V_{\omega_a}(x)\right)^2 - g_f(V_{\omega_a}(x)) \right] \right) \tag{11}$$

Importance weights can be computed based on the fact that the bound in Eq. 10 is tight for $T^*(x) = f'\left(\frac{p(x)}{q(x)}\right)$ where $f(u) = (u-1)^2$. We can therefore, approximate the desired importance weights as described in Eq. 3 as $w_{a,n} = \frac{g_f(V_{\omega a}(x_{a,n}))}{2} + 1$ for all $a \in (1, ...,A)$ and $n \in (1, ...,N_a)$. Ultimately, the ATE can be estimated between any two treatment arms according to Eq. 7. For example, the ATE between arms 1 and 2 could be estimated as $\widehat{ATE} = \sum_{n=1}^{N_1} [w_{1,n} Y_{1,n}] - \sum_{n=1}^{N_2} [w_{2,n} Y_{2,n}]$.

2.5. Practical considerations

In the original GAN and f-GAN formulations the gradients for the generator is replaced with a related gradient that significantly speeds convergence of the model. Because our objective is minimization of the true χ^2-divergence rather than perfect distributional matching, we do not employ this loss function trick but instead apply the gradient as derived from the loss function in Eq. 11.

Although it is the case that the domain of the Fenchel conjugate for the χ^2-divergence is $\mathbb{R}$, we constrained it to $t \geqslant -2$ which produces valid likelihood ratios.

Gradient descent-based optimization of GANs is a notedly difficult task [30,1,11]. Though many methods are proposed to stabilize training, we have found it sufficient to employ a set of algorithmic heuristics: (i) standardization of our data by the joint mean and variance over all A populations prior to training; (ii) periodically re-centering the distribution of each discriminator to a noisy estimate of the mean of the generator distribution. This re-centering is accomplished by setting the value of a vector that is added to the input of the discriminators.

The approach for minibatch stochastic gradient descent for the cGAN is shown in Algorithm 1. The objective function F (Eq. 11) is optimized by minimizing with respect to the parameters θ of the generator and maximizing with respect to the parameters $\omega_{1:A}$ of the discriminators.

2.6. Related work

Causal inference with observational data has a rich literature that cuts across many disciplines [44,37,38,33] including machine learning [18,21,41,34,40]. More specifically there have been several approaches to applying adversarial networks for counterfactual inference [21,45]. However, most existing methods for counterfactual inference are not directly comparable to the cGAN, as we aim to identify the most appropriate counterfactual distribution given the available data and maximize feature balance whereas most methods evaluate ATE estimation or individual treatment effect (ITE) estimation directly.

In contrast to representational learning approaches and some GAN approaches, our approach does not rely on a predefined outcome to identify matched cohorts. The approach outlined in [21] is the most similar in spirit to our approach but differs in that our objective directly minimizes the variance of expectations that might be used in ATE estimation, whereas [21] minimizes a bound on the variance of the

average treatment effect on the treated. As a result of this difference, unlike comparator methods, the cGAN requires neither regularization nor constrained optimization over weights.

3. Experiments

To evaluate the cGAN, including its utility in practice, we present results of a simulation and applications to real-world medical data.

3.1. Simulation

To evaluate the cGAN when the ground truth is known, we applied the model on simulated data of two populations/treatment arms, $A = 2$. Each population was comprised of two subpopulations. Each subpopulation contained 10 features, drawn from a randomly generated multivariate normal distribution with a normal-Wishart prior distribution. Population 1 was composed of an equal number of samples (N = 1000) from subpopulation A and subpopulation B; and Population 2 was composed of an equal number of samples from subpopulation A and subpopulation C (N = 2000). By construction, subpopulation A is a latent population associated with a natural experiment, since it is part of both Population 1 and 2.

Because our simulation deliberately constructs populations from a shared subpopulation distribution (A), we would expect points generated from this subpopulation to have higher weights. Intuitively, the variance of importance sampling estimates should be small for both treatment groups ($a = 1$ and $a = 2$) if the learned target distribution, P_0 is one that overlaps both populations maximally while excluding density unique to one group.

To better demonstrate how the cGAN supports counterfactual reasoning, we have additionally conducted an analysis of the average treatment effect (ATE) for our experiment with simulated data. We simulated a continuous outcome according to the subpopulation of origin – Pop 1A ~ Gaussian (60, 1); Pop 1B ~ Gaussian (40, 1); Pop 2A ~ Gaussian (-10, 1); Pop 2C ~ Gaussian (10, 1). Under this outcome function, the estimate of average treatment effect (ATE) under the mixture distribution (of Pop 1 and Pop 2) is 50. When estimating the ATE under the overlapping subpopulation distribution – those from Pop 1A and Pop 2A – the ATE is 70.

In this scenario, the 'treatment' of interest is the population (Pop 1 or Pop 2). According to counterfactual theory, unbiased ATEs arise in the presence of strong ignorability that is exemplified by distributional equality in the features. Given the structure of our simulation, strong ignorability between Pop 1 and Pop 2 is only upheld among those units from subpopulation A. Without this consideration, the ATE over the mixture distribution (Pop 1 vs Pop 2) will be confounded by units from subpopulation B and subpopulation C that differ on both the population and their features.

We applied weights from the cGAN and comparators to the simulated outcomes to assess the ability of the weighting methods to estimate one of the two ATEs. In addition, we also calculated the effective sample size (ESS), n_{eff}, using the Kish Method [26]. The ESS may be used to determine the quality of a Monte Carlo approximations of importance sampling. The calculation of n_{eff} can be found in the equation below, wherein w are the weights.

$$n_{eff} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

To investigate (i) feature-balancing weights, (ii) the biasedness of ATE, and (iii) the ESS, a variety of comparator methods and the cGAN were implemented and compared to the unweighted cohort. They include IPW, clipped-IPW in which propensity scores greater than 90th percentile and less than 10th percentile are assigned to the values of the percentiles at 90th and 10th, respectively [3]; binary regression propensity score; generalized boosted modeling of propensity scores [29]; covariate-balancing propensity scores [17]; non-parametric covariate-balancing propensity scores [8]; entropy balancing weights [12]; empirical balancing calibration weights [2]; and optimization-based weights [24].

To better understand how simulation parameters affect cGAN and comparator performance on ATE and ESS, we have additionally implemented a sensitivity analysis. This sensitivity analysis explores how combinations of (i) the per-arm sample size (N); (ii) the unbiased average treatment effect that exists in the truly counterfactual populations ('true' ATE); and (iii) the size of the truly counterfactual populations as a proportion of the total population (overlap) effect the outcome measures. In addition to the simulation parameters outlines above – which outlines a per-arm population size of 2000 in which the size of the truly counterfactual populations is 0.5 (50%) of the population, and an unbiased, 'true' ATE of 50 – simulations were replicated for all combinations of N = [2000, 4000, 8000], overlap = [0.1, 0.5, 0.9] and a 'true' ATE = [400, 70, 0.2]. This range for the sensitivity analysis represents the breadth of values that may be present in these parameters. To conduct this simulation and sensitivity analysis, parameters that the distribution of each subpopulations' features were randomly drawn according to all combinations of N and overlap, to create nine, unique populations for the sensitivity analysis. The cGAN and comparator methods were trained on each of these populations to learn feature-balancing weights. From each population, treatment effects for units from subpopulation A were generated according to each of the 'true' ATEs. Units that were not from subpopulation A had different treatment effects from subpopulation A, highlighting the possibility of heterogeneity of treatment effect in real-world scenarios. ATEs was calculated for all combinations of N, overlap, and 'true' ATE. ESS was calculated for all combinations of N and overlap, as this metric is independent of 'true' ATE.

To train the cGAN for the simulation and sensitivity analyses, the model was run for 200,000 training iterations at a learning rate of $1e^{-5}$.

3.2. Application to clinical data

We additionally applied the cGAN to an experiments using real-world clinical data from a large, academic medical center. For this experiment, we constructed the treatment and comparator cohorts according to the protocol and indication of a published randomized clinical trial. The experiment compares sitagliptin and glimepiride in elderly patients with Type II Diabetes Mellitus (N = 144 per arm) [13]. We present the 37 most frequent clinical measurements from the electronic health record.

We evaluate the ability of the cGAN to improve feature balance by comparing the Absolute Standardized Difference of Means (ASDM) between the treatment and comparator cohorts under different weighting methods. the ASDM is a popular method of assessing cohort similarity, with a lower metric corresponding to improved feature balance. The ASDM is presented for the cGAN and the comparator weighting methods mentioned in the Simulation Section (3.1). To train the cGAN for the Application to Clinical Data, the model was run for 2,000,000 training iterations at a learning rate of $1e^{-6}$. This learning rate is slightly smaller than that used in the simulation. It is a conservative method to prevent against mode-collapse in this higher-dimensional setting.

4. Results

4.1. Simulation

The results of our simulation is summarized in Figs. 2. In the left hand-side of the Figure, the columns show the marginals of three pairs of continuous features. Row (i) shows the raw data, colored by which population units were drawn from. Row (ii) shows the same data as above, but coloring by subpopulation to highlight the overlapping

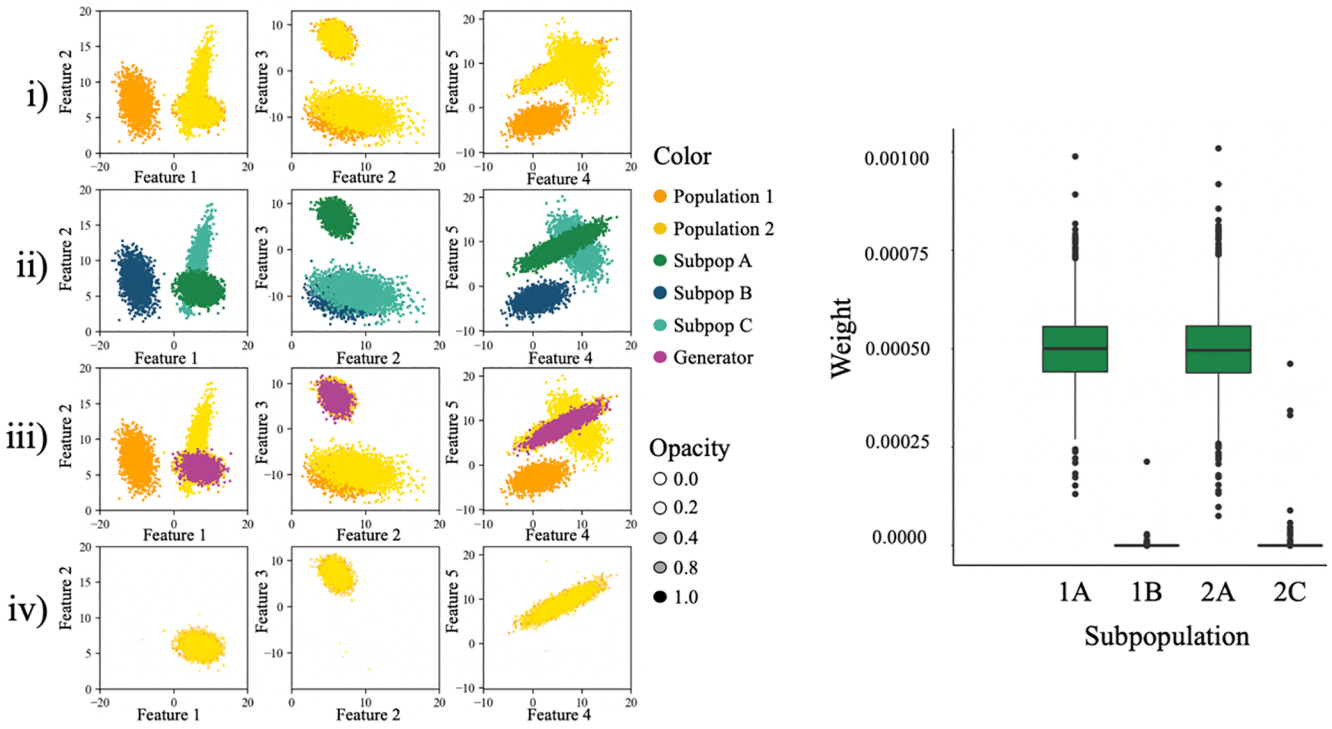

Fig. 2. Simulation Results. *Left:* Select features (i) by population of origin; (ii) with subpopulation A highlighted; (iii) samples from the generator; (iv) opacity adjusted by weight. *Right:* Weights by subpopulation.

Table 1
Results of Simulation. The average treatment effect and effective sample size (ESS) after application of weighting methods from the Counterfactual χ-GAN and comparators.

Weighting Method	ATE	ESS
unweighted	50.03	8000
IPW	92.00	6551
clipped IPW	87.24	6997
binary regression PS	92.00	6551
generalized boosted modeling PS	84.51	7207
covariate balancing PS	91.83	6686
non-parametric covariate balancing PS	37.65	11
entropy balancing	104.13	65
empirical balancing calibration weights	52.06	65
optimization-based weights	52.07	114
cGAN	**70.01**	**3870**

distribution. Row (iii) shows a set of samples from the generator after training colored in blue. Row (iv) depicts the original data from Row (i) with the opacity of data points reflecting the importance weights. The right-hand side of the Figure shows the distribution of weights by subpopulation. Note that, in both Populations 1 and 2, the mean weights of units from subpopulation A have weights near $5x10^{-4}$, which is the uniform weight when 2000 units are in each population. Units from other subpopulations have near negligible weights, and would not meaningfully contribute to expectations in Eq. 7.

In the left-most figure, as you move down any column of feature pairs, it is apparent that points from the overlapping subpopulation A are both captured by the generator and assigned higher weights. This is confirmed by plotting the weights of data points by subpopulation (right-hand side of 2). Weights from subpopulations 1A and 2A are substantially higher than those from subpopulations 1B and 2C.

The results of this simulation further demonstrate that the ATE estimate from cGAN-weighted data is less biased than estimates from other weighting methods, given their respective targets (Table 1). By construction, the causal effect of the comparable subpopulations is 70. cGAN-weighted data produced an ATE of 70.01. We see similarly good performance when inspecting the ESS. The cGAN has an ESS of 3870. Given that there are 4000 units that are comparable across the two arms (each subpopulation contains 2000 units), this is an appropriate estimate.

The results of the sensitivity analyses can only be found in the Supplementary Materials. The results of this analysis show that superiority of cGAN performance over comparator methods persists across all settings of the simulations parameters. Across all combinations of per-arm sample size, overlap, and 'true ATE, the cGAN consistently produced the least biased estimate of ATE and yielded the maximally appropriate ESS given the parameters.

4.2. Application to clinical data

The ASDM for the clinical cohorts is presented in Fig. 3. These findings are summarized by the mean ASDM over all features, under the varying weighting methods in Table 2. cGAN improved mean ASDM from the unweighted cohort and improved feature balance the most among all evaluated methods. Under cGAN-weighting, some features show worsening ASDM after weighting is applied. We hypothesize that this may be due to incomplete training, and may be alleviated with more iterations. Note that this task is particularly challenging due to the high dimensionality of the data and small study size.

The results of this experiment can be found in Fig. 3 and Table 2. They demonstrate that cGAN-weighting achieves better feature balance than comparator methods.

5. Discussion

In this paper, we introduce the Counterfactual χ-GAN. It is a deep generative model for feature balance that minimizes the variance of importance sampling estimates of treatment effects. We leverage the *f*-GAN framework for estimating the χ^2-divergence and likelihood ratios necessary for achieving this.

The experiments presented here suggest that cGAN is an effective

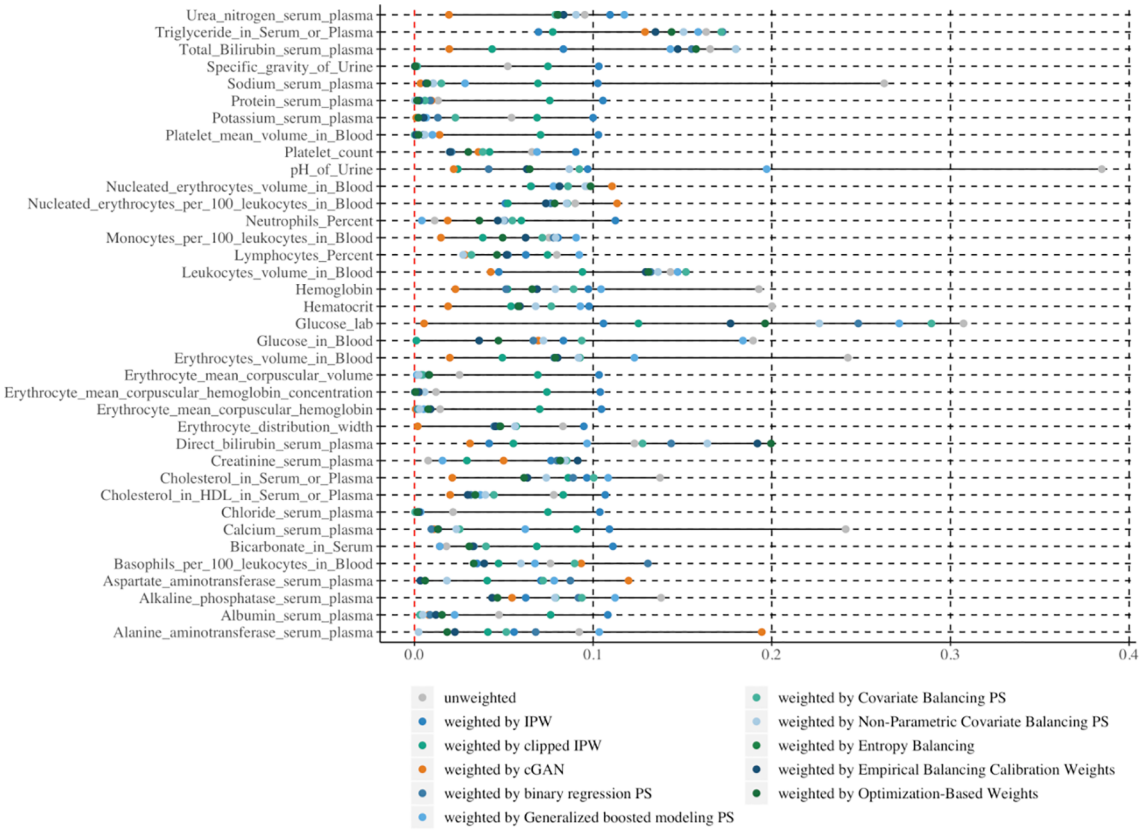

Fig. 3. Absolute standardized difference of the means (ASDM) of real-world clinical features after application weighting methods from the Counterfactual χ-GAN and comparators.

Table 2
Results of Application to Clinical Data. Absolute standardized difference of the means (ASDM) of real-world clinical features after application weighting methods from the Counterfactual χ-GAN and comparators.

Weighting Method	ASDM
unweighted	0.1103
IPW	0.0876
clipped IPW	0.0631
binary regression PS	0.0625
generalized boosted modeling PS	0.0749
covariate balancing PS	0.0681
non-parametric covariate balancing PS	0.0596
entropy balancing	0.0524
empirical balancing calibration weights	0.0524
optimization-based weights	0.0536
cGAN	**0.0364**

method of learning feature balancing weights to support counterfactual inference. If we assume that all potentially confounding variables are observed, the superiority of cGAN in learning balancing weights, suggests that ATE estimates borne from cGAN-weighted cohorts would be less biased than those estimates generated from traditional weighting methods. As such, the cGAN may provide a reliable means to better understand the safety and effectiveness of interventions, when a randomized trial is not feasible. This may get high-quality interventions to patients faster, thereby improving patient health.

The application of the model to real-world EHR data, demonstrates that this method could provide an alternative means to causal estimation from observational data when the assumptions of no unobserved confounding, positivity, and SUTVA are met. This finding could have been empirically verified through a comparison of the cGAN-weighted ATE with the gold standard, which is randomized trial in which strong ignorability should be upheld. The effect of randomization in the RCT and cGAN-weighting only addresses the differences that exist between the treatment and comparator cohorts within a single experimental setting. However, we found that the experimental setting from the RCT were markedly different from the observational data, notably in the distribution of gender and race. If the variables with differing distributions elicit a heterogeneity of treatment effect, then the effect estimates between these two data sources would never be comparable.

Our experiments suggest that the flexibility of our framework produces improved feature balance relevant for valid causal estimates. This method does, however, come with limitations. GANs are well known for their instability and lack of objective measures for convergence. This work shares those limitations. In future work, we will explore an extension of the cGAN that overcome the many current limitations of GANs. To make the model of use to the informatics community, we may extend upon our model and develop a new method that still minimizes the χ-divergence but does so in a way that is (i) rapid, (ii) allows monitoring of convergence, (iii) that is more robust to hyperparameter settings, and (iv) handles heterogeneous and missing data seamlessly.

CRediT authorship contribution statement

AA contributed Investigation, Methodology, Software, Validation, Visualization, and Writing - original draft. AP contributed Conceptualization, Methodology, Project administration, Supervision, and Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by grants R01LM009886-10 and T15LM007079 from The National Library of Medicine.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2020.103515.

References

[1] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, 2017.
[2] K.C.G. Chan, S.C.P. Yam, Z. Zhang, Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting. Technical report, University of Washington, 2016.
[3] S.R. Cole, M.A. Hernán, Constructing inverse probability weights for marginal structural models, Am. J. Epidemiol. 168 (6) (2008) 656–664.
[4] O.M. Dekkers, E. von Elm, A. Algra, J.A. Romijn, J.P. Vandenbroucke, How to assess the external validity of therapeutic trials: A conceptual approach, Int. J. Epidemiol. 39 (1) (2010) 89–94.
[5] A.B. Dieng, D. Tran, R. Ranganath, J. Paisley, D.M. Blei, Variational Inference via χUpper Bound Minimization, in: NIPS, 2017.
[6] J.A. Dimasi, R&D Cost Study Briefing Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs, 2014.
[7] A. Doucet, N. Freitas, N. Gordon, An introduction to sequential Monte Carlo methods, in: Sequential Monte Carlo Methods in Practice, Springer, New York, New York, NY, 2001, pp. 3–14.
[8] C. Fong, C. Hazlett, K. Imai, Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements, Ann. Appl. Stat. 12 (1) (2018) 156–177.
[9] T.A. Glass, S.N. Goodman, M.A. Hernán, J.M. Samet, Causal inference in public health, Ann. Rev. Public Health 34 (2013) 61–75.
[10] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Schölkopf, J. Candela, M. Sugiyama, A. Schwaighofer, N. Lawrence, Covariate shift by kernel mean matching, Dataset Shift Mach. Learn. 131–160 (2009) (2009).
[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved Training of Wasserstein GANs, in: NIPS, 2017.
[12] J. Hainmueller, Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies, Polit. Anal. 16 (2011) 25–46.
[13] P. Hartley, Y. Shentu, P. Betz-Schiff, G.T. Golm, C.M. Sisk, S.S. Engel, R.R. Shankar, Efficacy and tolerability of sitagliptin compared with glimepiride in elderly patients with type 2 diabetes mellitus and inadequate glycemic control: a randomized, double-blind, Non-Inferiority Trial. Drugs Aging 32 (6) (2015) 469–476.
[14] C. Hazlett, Kernel Balancing: A Flexible Non-Parametric Weighting Procedure for Estimating Causal Effects, SSRN, 2016.
[15] J.K. Hellerstein, G.W. Imbens, Imposing moment restrictions from auxiliary data by weighting, Rev. Econ. Stat. 81 (1999) 1–14.
[16] P.W. Holland, Statistics and Causal Inference, JASA 81 (396) (1986) 945–960.
[17] K. Imai, M. Ratkovic, Covariate Balancing Propensity Score, Technical Report, Harvard University, 2013.
[18] F.D. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, 2016.
[19] N. Kallus, Causal inference by minimizing the dual norm of bias: Kernel matching & weighting estimators for causal effects, in: CEUR Workshop Proceedings, vol. 1792, 2016, pp. 18–28.
[20] N. Kallus, A framework for optimal matching for causal inference, in: AISTATS, 2017, pp. 372–381.
[21] N. Kallus, Deepmatch: Balancing deep covariate representations for causal inference using adversarial training, 2018a.
[22] N. Kallus, Optimal a priori balance in the design of controlled experiments, J.R. Statist. Soc. B 80 (1) (2018) 85–112.
[23] J.D.Y. Kang, J.L. Schafer, Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data, Stat. Sci. 22 (4) (2007) 523–539.
[24] L. Keele, J. Zubizarreta, Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the School Voucher System in Chile, 2014. arXiv preprint arXiv:1409.8597 1–37.
[25] O. Kempthorne, The randomizaton theory of experimental inference, J. Am. Stat. 50 (1955) 946–967.
[26] L. Kish, Survey Sampling, Survey Sampling, Wiley, New York, 1965(Chapter 14).
[27] A.S.S. Leger, Statistical Models in Epidemiology vol. 48, Oxford University Press, 1994.
[28] F. Li, K.L. Morgan, A.M. Zaslavsky, Balancing covariates via propensity score weighting, J. Am. Stat. Assoc. 113 (521) (2018) 390–400.
[29] D.F. McCaffrey, G. Ridgeway, A.R. Morral, Propensity score estimation with boosted regression for evaluating causal effects in observational studies, Psychol. Methods 9 (4) (2004) 403–425.
[30] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge?, 2018.
[31] M.E. Muller, Review: J.M. Hammersley, D.C. Handscomb, Monte Carlo Methods; Yu. A. Shreider, Methods of Statistical Testing/Monte Carlo Method, vol. 37, Springer Netherlands, 1966.
[32] S. Nowozin, B. Cseke, R. Tomioka, f-GAN: training generative neural samplers using variational divergence minimization, NIPS, 2016, pp. 271–279.
[33] J. Pearl, Causality, Cambridge University Press, Cambridge, England, 2000.
[34] M. Ratkovic, Balancing Within the Margin: Causal Effect Estimation with Support Vector Machines, Department of Politics, Princeton University, Princeton, NJ, 2014.
[35] P.R. Rosenbaum, D.B. Rubin, Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome, 1983.
[36] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. 27 (3) (1956) 832–837.
[37] D.B. Rubin, The use of matched sampling and regression adjustment to remove bias in observational studies, Biometrics 29 (1) (1973) 185.
[38] D.B. Rubin, Estimating causal effects of treatments in randomized and non-randomized studies, J. Educ. Psychol. 66 (5) (1974) 688–701.
[39] D. Rubin, Randomization analysis of experimental data: The Fisher randomization test comment, JASA 75 (371) (1980) 591–593.
[40] P. Schwab, L. Linhardt, W. Karlen, Perfect match: A simple method for learning representations for counterfactual inference with neural networks, 2018.
[41] U. Shalit, F.D. Johansson, D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, in: ICML, 2017.
[42] C. Tao, L. Chen, R. Henao, J. Feng, L.C. Duke, Chi-square Generative Adversarial Network, 2018.
[43] F. Thoemmes, A.D. Ong, A primer on inverse probability of treatment weighting and marginal structural models, Emerg. Adulthood 4 (1) (2016) 40–59.
[44] M. Thrusfield, Observational studies, Veterinary Epidemiology, fourth ed., Springer, 2017, pp. 319–338.
[45] J. Yoon, J. Jordon, M. van der Schaar, GANITE: estimation of individualized treatment effects using generative adversarial nets, in: ICLR, 2018.