Loiseau *et al.*
BMC Medical Research Methodology (2022) 22:335
https://doi.org/10.1186/s12874-022-01799-z

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning

Nicolas Loiseau[*†], Paul Trichelair[†], Maxime He, Mathieu Andreux, Mikhail Zaslavskiy, Gilles Wainrib and Michael G. B. Blum

Abstract

Background: An external control arm is a cohort of control patients that are collected from data external to a single-arm trial. To provide an unbiased estimation of efficacy, the clinical profiles of patients from single and external arms should be aligned, typically using propensity score approaches. There are alternative approaches to infer efficacy based on comparisons between outcomes of single-arm patients and machine-learning predictions of control patient outcomes. These methods include G-computation and Doubly Debiased Machine Learning (DDML) and their evaluation for External Control Arms (ECA) analysis is insufficient.

Methods: We consider both numerical simulations and a trial replication procedure to evaluate the different statistical approaches: propensity score matching, Inverse Probability of Treatment Weighting (IPTW), G-computation, and DDML. The replication study relies on five type 2 diabetes randomized clinical trials granted by the Yale University Open Data Access (YODA) project. From the pool of five trials, observational experiments are artificially built by replacing a control arm from one trial by an arm originating from another trial and containing similarly-treated patients.

Results: Among the different statistical approaches, numerical simulations show that DDML has the smallest bias followed by G-computation. In terms of mean squared error, G-computation usually minimizes mean squared error. Compared to other methods, DDML has varying Mean Squared Error performances that improves with increasing sample sizes. For hypothesis testing, all methods control type I error and DDML is the most conservative. G-computation is the best method in terms of statistical power, and DDML has comparable power at $n = 1000$ but inferior ones for smaller sample sizes. The replication procedure also indicates that G-computation minimizes mean squared error whereas DDML has intermediate performances in between G-computation and propensity score approaches. The confidence intervals of G-computation are the narrowest whereas confidence intervals obtained with DDML are the widest for small sample sizes, which confirms its conservative nature.

Conclusions: For external control arm analyses, methods based on outcome prediction models can reduce estimation error and increase statistical power compared to propensity score approaches.

Keywords: Observational study, Average treatment effect, Propensity score, Confounding variables, Replication study, Counterfactual, Doubly robust

[†]Nicolas Loiseau and Paul Trichelair contributed equally to this work.

*Correspondence: nicolas.loiseau@owkin.com

Owkin France, Paris, France

Background

There is an increasing interest in using external control arms (ECA) as a source of evidence to assess treatment efficacy. An ECA consists of a cohort of patients that serve as controls to an intervention arm from a clinical trial, and these control patients are collected from data sources external to the single-arm trial [1, 2]. After running a single-arm phase 2 study, usage of ECA is relevant to reduce false positive rates [3]. ECAs are also relevant to supplement randomized trials when randomization is unethical or when it is difficult to recruit patients, typically for rare diseases or in precision oncology where recruitment relies on biomarkers [4].

However, causal inference in non-randomized studies such as ECA is prone to confounding bias [5, 6]. Without randomization, estimation of treatment effect can be biased partly because of differences between the characteristics of patients in the two arms. Methods based on propensity scores are well established to account for confounding factors [7–10]. Propensity scores relies on the exposure model that provides a mapping between patient characteristics and the probability to be in the external arm. As an alternative, there are several methods that require prediction of clinical outcomes based on covariates and on treatment [11, 12]. In epidemiology, G-computation is such an alternative, and it is based on the *counterfactual framework* in which we posit that we can predict a patient outcome if the patient would have been enrolled in the control arm instead of the experimental one or vice-versa, making the inference of a causal effect theoretically possible [11]. With the advent of causal inference in machine learning, the counterfactual framework has been re-investigated and new methods were proposed including doubly debiased machine learning [12], which addresses bias of machine learning estimators. Here, we consider both synthetic simulations and data of clinical trials to evaluate statistical properties of both propensity score and outcome prediction methods. Evaluated methods seek to estimate the average treatment effect on the treated (ATT), which is defined as the benefit of the investigated treatment when averaged over the characteristics of the individuals originating from the intervention arm of the clinical trial.

The first class of statistical methods relies on propensity scores that are computed after learning an exposure model e, which relates individual covariates to the probability to lie in the experimental arm. Exposure model can be estimated using a logistic regression. Treatment effect is then estimated using patients matching and/or weighting, such as the distribution of the propensity scores should be the same in both arms. Rosenbaum an Rubin [13] showed that if positivity and conditional ignorability hold, then conditioning on the propensity score allows to obtain unbiased estimates of average treatment effects [14]. Conditional ignorability means that there are no unmeasured confounders. Mathematically, it states that given a set of covariates X, treatment assignment T is independent of the potential outcomes (Y^0, Y^1) that would be realized when the treatment T is equal to 0 (control) and 1 (investigated treatment). The second assumption is positivity and it assumes that $0 < P(T = 1|X) < 1$, for all values of X, which means that every subject has a nonzero probability to receive the control treatment and the investigated treatment. If the exposure model is misspecified, potentially because parametric assumptions of logistic regression are not valid, then estimators of treatment effect might be biased [15].

The second class of methods, outcome prediction methods, relies on the outcome model μ_0, sometimes named Q-model, which is the conditional expectation of the clinical outcome based on covariates X [11]. Because we focus on the estimation of the average treatment effect on the treated (ATT), the nuisance function μ_0 corresponds to the expected outcome for a patient enrolled in the control arm (see Section 2). By contrast, estimation of the average treatment effect (ATE) would have required outcome prediction as function of both the treatment and the covariates, which is the standard definition of the Q model [11]. Fitting the Q model can be done with flexible machine learning models such as boosted trees or neural networks [16, 17]. Machine learning models can be trained using regularization to limit overfitting. However, while reducing variance of estimators, regularization can bias estimation of outcome model that can in turn bias estimation of treatment effect [12]. Doubly debiased machine learning (DDML) is related to G-computation but it further accounts for the possible bias of machine learning outcome models [12]. DDML requires to estimate both the exposure model e and the outcome model μ_0, and flexible models can be fitted to infer both e and μ_0, which are considered as nuisance parameters [18]. DDML is an instance of a doubly robust estimator because it requires that only one of the exposure and outcome models need be correctly specified in order to obtain an unbiased estimator of treatment effect. To provide unbiased estimation of treatment effect, DDML relies on Neyman orthogonal scores and on cross fitting, which is a sample splitting approach [12].

There is a lack of studies based on clinical trial data that compares propensity score approaches and methods based on outcome modelling. Propensity score matching and weighting are two common methods used to provide evidence of drug effectiveness and we seek to evaluate to what extent statistical analysis can be improved with outcome based modelling. Numerical simulations suggest that G-computation reduces bias and variance of causal inference estimate compared to propensity-score approaches [19, 20]. Another simulation study finds that DDML was among the top performers methods to estimate average treatment effect [21]. However, comparisons based on actual trial data are insufficient. Here we consider an internal replication framework for evaluation of causal inference methods [22]. It is based on comparisons between randomized studies that provide ground truths for treatment effect and artificial non-randomized studies consisting of the grouping of the experimental arm and of the standard-of-care arm, which are derived from two different clinical trials [23]. An internal replication framework was used for instance to demonstrate that propensity score matching is highly sensitive to baseline covariates included in the exposure model [24]. Internal replication framework are not the only setting to compare results from RCT and from observational data. Several studies compared results obtained from observational data to the conclusions obtained from randomized experiments, which are considered as ground truth [25–28]. However, heterogeneity of treatment effect can explain the difference of efficacy measured in a RCT and observational setting [29, 30]. By contrast, there is no expected difference of treatment effect (ATT) in internal replication studies when comparing efficacy obtained from randomized and non-randomized experiment [22]. Our internal replication study is based on data from the YODA project, which includes a pool of type 2 diabetes randomized clinical trials sharing arms with the same treatment delivered to patients (Canagliflozin) [31, 32].

Methods

Average treatment effect on the treated (ATT)

Generally, the primary quantity of interest in interventional clinical trials is the efficacy of an investigated treatment compared to another standard of care or placebo treatment. Formally, from the study cohort comprising of two groups, each exposed to a different treatment T (0 for control, 1 for experimental treatment), the target is to infer the average treatment effect on the treated (ATT). The ATT corresponds to the difference between the outcome of a patient treated with the experimental drug and a control patient when averaging over baseline clinical attributes X of patients belonging to the experimental treatment arm. Using the formalism of potential outcomes, the ATT is defined as [33]

$$\mathrm{ATT} = \mathbb{E}\left[Y^1 - Y^0 | T = 1\right],$$

where Y^0 (respectively Y^1) is the potential outcome for a unit that undergoes treatment 0 (respectively 1). The observed outcome Y can be expressed as

$$Y = Y^1 T + Y^0 (1 - T).$$

For a given patient, only one of the two potential outcomes is realized and observed, the other is named a counterfactual outcome. The ATT estimand is different from the average treatment effect that is obtained as

$$\mathrm{ATE} = \mathbb{E}\left[Y^1 - Y^0\right].$$

If either the propensity score $e(X) = \mathbb{E}[T|X]$ is constant (randomization) or the Conditional Average Treatment Effect $\mathrm{CATE} = \mathbb{E}\left[Y^1 - Y^0 | X\right]$ is constant (no heterogeneity), then ATT and ATE are equal. In the following, we will also denote by $\mu_0 = \mathbb{E}\left[Y^0 | X\right]$, the conditional expectation of the outcome for patients in the control arm.

Estimators of the average treatment effect of the treated

The problem of causal inference for external control arm analysis revolves around the two populations' prognosis characteristics not being of equal distribution in the two arms. A solution to balance populations' characteristics is to reweight or choose units such that the two resulting virtual populations match as closely as possible. To balance populations, the exposure model $e(\cdot)$ should be estimated when considering propensity score matching (PSM) and Inverse Probability of Treatment Weighting (IPTW). The PSM estimator selects matched units in each group whereas IPTW re-weights units based on functions of the propensity score, which leads to the following estimator [34, 35]

$$A\hat{T}T_{IPTW} = \frac{1}{n_1} \sum_{i=1}^{n} Y_i \left(T_i - \frac{\hat{e}(X_i)(1 - T_i)}{1 - \hat{e}(X_i)} \right),$$

where X_i, Y_i, T_i are the covariates, outcome, and treatment for the i^{th} individual, $1, \ldots, n_1$ are the indices of the individuals in the experimental arm, $n_1 + 1, \ldots, n$ are the indices of the individuals in the external arm, n, n_1 are the sample sizes for the whole sample and the experimental arm only, and $\hat{e}$ is an estimator of the exposure model.

For propensity score matching, we consider a greedy nearest neighbor algorithm without replacement with a caliper as an hyper-parameter. The algorithm consists in selecting the pairs of one treated patient and one control patient with the closest propensity score values. The algorithm iterates until there is no patient left in the treated arm or if the propensity score distances of the non-selected individuals are greater than a caliper of 0.25. The patients that have not been selected are discarded from the downstream efficacy analysis.

The first estimator based on outcome prediction we consider is the G-computation estimator [19, 36]. G-computation does not rely on estimation of the propensity score but on the the conditional expectation of the outcome μ_0. For each treated patient defined by his covariates X and outcome Y, we can predict a control counterfactual outcome $\hat{\mu}_0(X)$, and the G-computation estimator is defined as the average over the experimental arm of the difference between the measured and counterfactual outcome

$$\hat{ATT}_{GC} = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - \hat{\mu}_0(X_i)), \tag{1}$$

where $\hat{\mu}_0$ is an estimator of the nuisance function.

Machine learning estimators can be biased in order to avoid overfitting and this is especially true when the dimension of the covariates X is large [37]. Doubly debiased machine learning (DDML) accounts for the bias of the G-computation estimator, which can result from the bias of a machine learning estimator, $\hat{\mu_0}$, for μ_0 [12]. A core principle of DDML is to consider a sample splitting approach to estimate and account for the bias of the machine learning estimator of the outcome model. The dataset is split into a training set and an auxiliary set. The training set is used to fit two machine learning models to learn the outcome and exposure models μ_0 and e. The ATT estimator is obtained by subtracting to the G-computation estimator evaluated on the auxiliary dataset an estimate of its bias

$$\hat{ATT}_{DDML} = \hat{ATT}_{GC} - \frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}}\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(1-T_i)(y_i - \hat{\mu}_0(X_i)), \tag{2}$$

where $\tilde{n}, \tilde{n_1}$ are the sample sizes for the whole auxiliary dataset and the control arm part of this dataset. Because the estimator depends on actual splitting, we consider an averaging procedure over multiple splits [12]. In the Appendix, we describe the averaging procedure and the estimation procedure for the variance.

Finally, we compute an unadjusted estimator that consists of the difference between the mean of the clinical outcomes Y in each arm. This estimator measures the level of bias that is expected when not accounting for confounding factors. If the data include confounding that may impact causal inference, the unadjusted estimator should be biased.

Variance, confidence intervals, and regularisation

To estimate variance and confidence intervals we consider non-parametric bootstrap for both the propensity score approach and G-computation. The bootstrap procedure is applied based on 300 replicates of the original dataset. For G-computation, bootstrap includes training of the functions μ_0 and e and for propensity score approaches, it includes only training of the propensity score function e. For DDML, we consider a sample-splitting approach [12], and the estimation procedure is detailed in the Appendix.

For all methods, we consider linear regression and logistic regression with all covariates to fit μ_0 and e. To train the propensity score model e, we consider ridge regression, and to train the outcome model μ_0, we consider lasso regression. For G-computation, regularisation parameters were learned using cross-validation. For DDML, regularisation parameters were learned using nested cross-validation because of the internal cross-validation procedure described in the Appendix. Machine learning operations were performed using the *Scikit-learn* Python library [38].

Synthetic simulations

We consider two scenarios of simulations to benchmark estimators. The first scenario assumes an homogeneous treatment effect and includes confounding factors because both the exposure and the outcome models are linear functions of several of the 20 simulated covariates. The second scenario further assumes an heterogeneous treatment effect by including interaction between treatment and covariates to model outcomes.

Experiments are based on synthetic data with a binary exposure T and 20 covariates X. The numbers of patients (including patients in both arms) of 250, 500 and 1000, were chosen to be in the same order of magnitude as external control arm analyses. The simulations rely on two scenarios differing by the potential outcomes (Y^0, Y^1) generation. For both scenarios, the exposure model is a linear function of 5 of the 20 covariates.

$$\text{logit}(\mathbb{E}[T|X]) = \frac{1}{\sqrt{5}}\sum_{j=1}^{5}\beta_j X^{(j)},$$

where $\beta_j \sim \mathcal{U}([-1,1])$, $X \sim \mathcal{N}(0,\Sigma)$ with Σ a random sparse symmetric definite positive matrix simulated for every simulated data using the scikit-learn function *make_sparse_spd_matrix* with an α value of 0.8, and where $X^{(j)}$ is the j^{th} element of the vector of covariates X. In the first scenario, the potential outcomes are sparse linear functions of the covariates and the treatment effect is homogeneous among the patients. To make it sparse, half of the variables are randomly sampled and the corresponding coefficient is set to zero,

$$y = f(X, \Omega) + \theta T + \epsilon, \text{with} f(X, \Omega) = \frac{1}{\sqrt{10}} \sum_{j=1}^{10} x^{\Omega(j)},$$

where $\epsilon \sim \mathcal{N}(0, 1)$, Ω is a random permutation of the covariate indices and $\theta \sim \mathcal{N}(0, 0.4)$ or $\theta = 0$ for the null hypothesis. We chose a variance of 0.4 because we have found in simulations that this value induces a level of confounding that biases the unadjusted estimator, and which can be handled with causal inference approaches.

The second scenario includes a term of interactions to model an heterogeneity of treatment. The outcome is obtained as follows :

$$y = (1 - T)f(X, \Omega_0) + Tf(X, \Omega_1) + \theta T + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$, Ω_1, Ω_2 are random permutations of the covariate indices, and θ is sampled such that $\text{ATT} \sim \mathcal{N}(0, 0.4)$ or $\text{ATT} = 0$.

To evaluate the estimators, the following metrics were considered : bias, mean absolute error (MAE), mean squared error (MSE), average confidence interval length measured by the variance of a matched Gaussian distribution, type I error and power.

Internal replication study

The internal replication study is based on data from five randomized clinical trials assessing the efficacy of Canagliflozin in patients with type 2 diabetes [39–43]. Access to the trials, shortly described in Table 1, was granted through the Yale University Open Data Access (YODA) Project [31, 32]. Experiments are restricted to the set of patients that share similar background therapy and inclusion/exclusion criteria in order to make causal inference valid because of the positivity assumption. A set of 40 baseline covariates were selected by a clinician and considered as confounding factors (see Appendix). The primary endpoint is change in HbA1c (glycated hemoglobin) between baseline and 12 weeks, which is available in all trials. Patients with missing outcome are not considered in the analysis.

From the pool of five trials, an observational setting is built by replacing a control arm in one trial by another trial arm composed of patients that were given the same treatment. This procedure is replicated by varying the trial of interest. Estimation obtained in the non-randomized setting can be compared to the treatment effect obtained in the well randomized setting.

We conduct two categories of internal replication studies. For each experiment, the experimental arm and the control arm are extracted from different trials. In the first category, the experimental and control treatments are the same. In this negative control setting, the treatment effect on the treated is null regardless of the underlying population [44]. The negative control study is based on 9 non-randomized comparisons. The ground truth of a null effect being known, the comparison between the estimators is performed using the following metrics: the mean absolute error (MAE), the mean squared error (MSE), the width of confidence intervals, and the coverage rates for the 95% confidence intervals.

Table 1 Description of the five type 2 diabetes clinical trials used for the internal replication study. We report only the trial-specific inclusion criteria

Trial	Nb. patients	Inclusion criteria	Arms	Background therapy
NCT01106625 [39]	469		Canagliflozin 300 Sitaglipin 100	Metformin and Sulphonylurea
NCT01137812 [40]	755		Canagliflozin 300 Canaglifozin 100 Placebo	Metformin and Sulphonylurea
NTC01106651 [41]	659	Age: 55 to 80 y.o.	Canagliflozin 300 Canaglifozin 100 Placebo	Metformin and Sulphonylurea (357 patients) Metformin (302 patients)
NCT01106677 [42]	1284		Canagliflozin 300 Canaglifozin 100 Sitaglipin 100 Placebo	Metformin
NCT00968812 [43]	1450	45≥BMI≥22	Canagliflozin 300 Canaglifozin 100 Glimepiride 100	Metformin

In the second category of experiments, the experimental and control treatments are different and an RCT estimate is available from one of the five trials listed in Table 1. In this RCT replication setting, a reference treatment effect and confidence intervals are available from the RCT but the true treatment effect is unknown. The RCT replication study is based on 19 non-randomized comparisons. Evaluation relies on previously proposed metrics [45]:

- Pseudo bias is defined as the difference between the randomized treatment effect estimation and the non-randomized estimation;
- Pseudo mean squared error is defined as the squared difference between the randomized effect estimation and the non-randomized estimation, averaged over the different combinations of trials;
- Estimate agreement measures the percentage of time when treatment effect estimated in the non-randomized setting lies within the 95% confidence interval of the randomized trial;
- Regulatory agreement is the percentage of time the cutoff $P < 0.05$ obtained with the non-randomized experiments agrees with the RCT result about $P < 0.05$.

Results

Synthetic simulations

Type I error rates and statistical power are evaluated with simulations using $P < 0.05$ as a decision cutoff. The unadjusted estimator has an inflated type I error ranging from $10 - 20\%$ when $n = 250$ to $30 - 40\%$ when $n = 1000$ showing that simulations include a confounding bias (Fig. 1). The statistical power obtained with the unadjusted estimator is of poor relevance because of its inflated type I error. All methods that adjust for confounding bias properly control type I error (Fig. 1). G-computation has a type I error of 5% whereas DDML is more conservative.

When considering G-computation, power is increased by $0 - 10\%$ when compared to propensity-score approaches (Fig. 1). By contrast, the power of DDML is not always larger than the ones of propensity-score approaches. The power of DDML is smaller that the ones obtained with propensity score approaches when

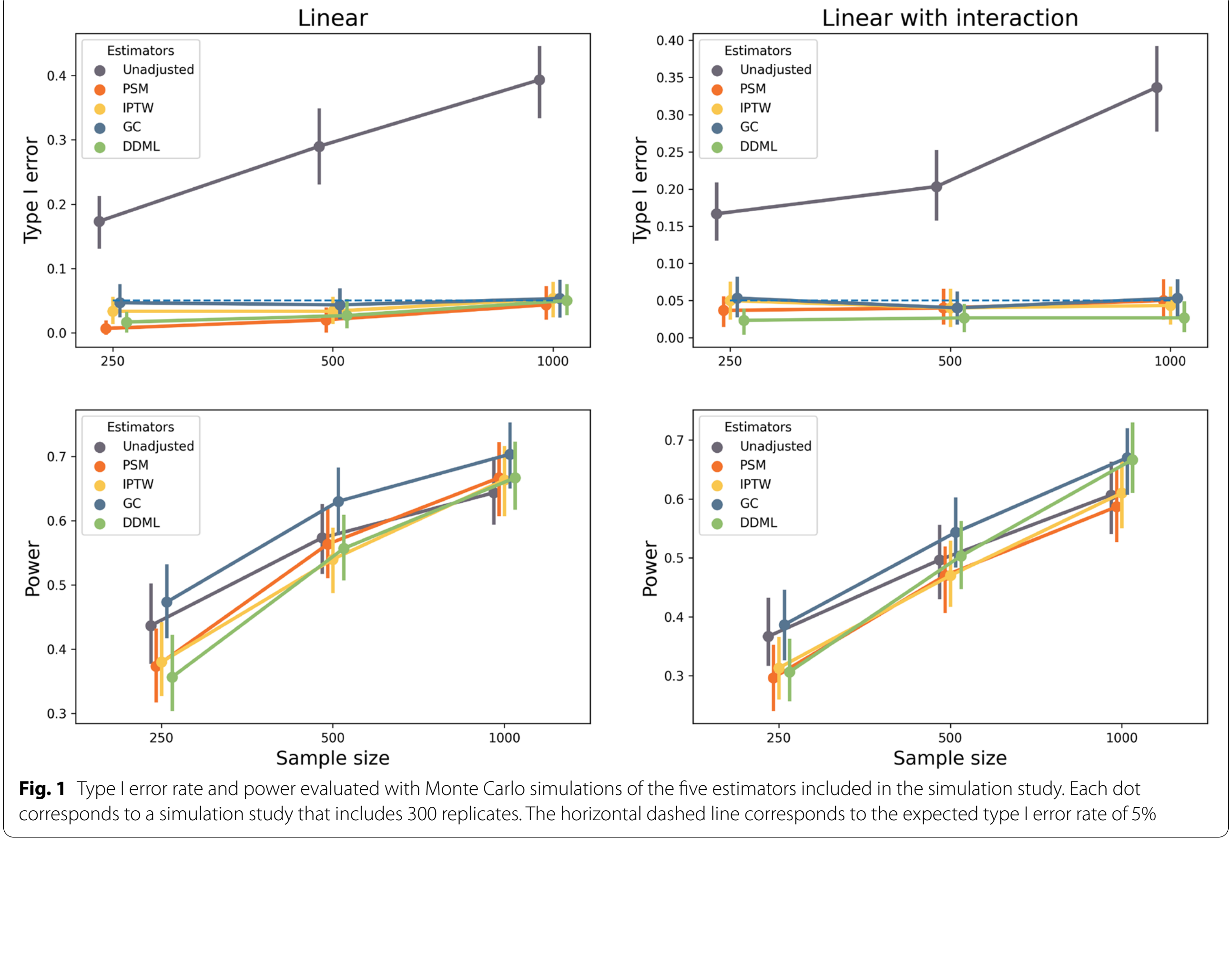

Fig. 1 Type I error rate and power evaluated with Monte Carlo simulations of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates. The horizontal dashed line corresponds to the expected type I error rate of 5%

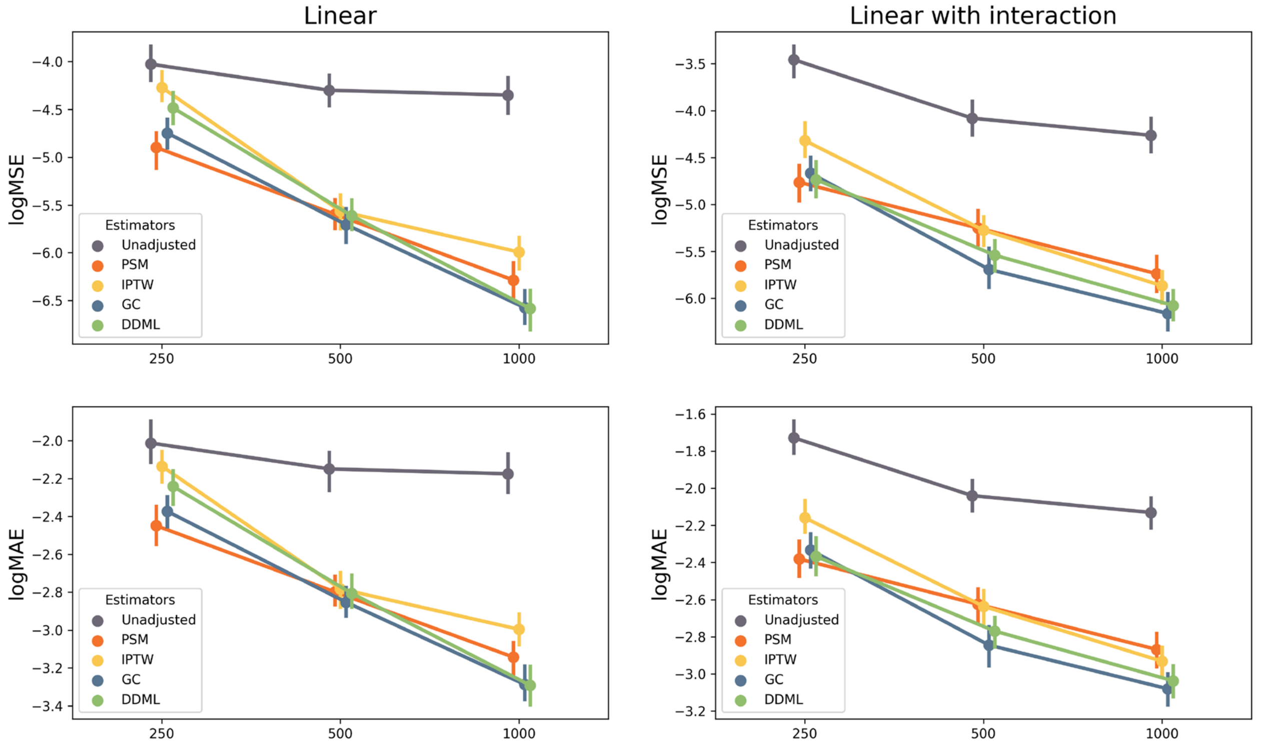

Fig. 2 Logarithm of the Mean Absolute Error (MAE) and Mean Squared Error (MSE) of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates

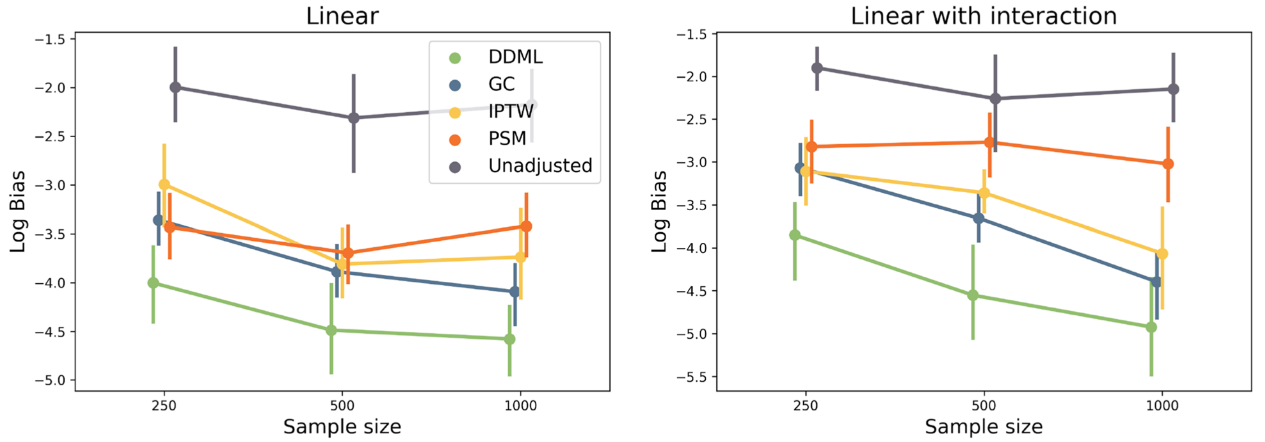

Fig. 3 Logarithm of the bias of the five estimators included in the simulation study. Each dot corresponds to a simulation study that includes 300 replicates

$n = 250$, of comparable values at $n = 500$, and larger when $n = 1000$. As expected, the power of each method increases with increasing sample size.

Statistical properties of the different estimators are also compared using Mean Absolute Error (MAE) and the Mean Squared Error (MSE) (Fig. 2). At $n = 250$, PSM generally achives the smallers errors. However, as for IPTW, their errors decrease more slowly as a function of sample size compared to outcome modelling methods. At $n = 1,000$, outcome modelling approaches achieve the smallest errors.

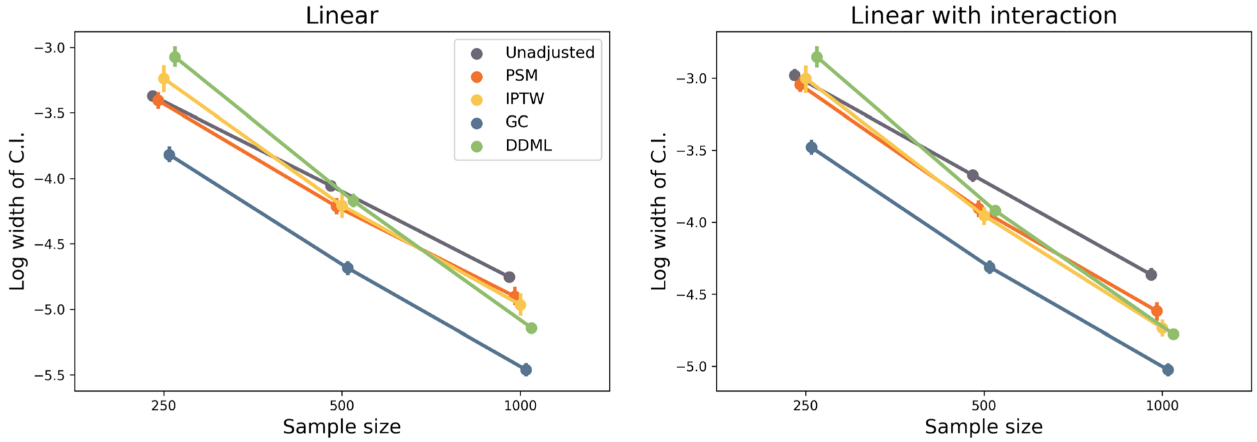

Fig. 4 Log width of the 95% Confidence Intervals (C.I) for the different methods. To measure the log width, we compute the logarithm of the variance of a Gaussian distribution which 95% C.I. would match the observed C.I. Each dot corresponds to a simulation study that includes 300 replicates

To have a finer look at the different properties of ATT estimators, we investigate their bias (Fig. 3). As expected by construction of the DDML estimator, its bias is inferior to the bias of G-computation. The bias of propensity score methods was larger than the ones of outcome prediction methods. The bias of outcome prediction method monotonically decreases as a function of sample size, which is not always the case of propensity-score methods.

We investigate the width of the confidence intervals by computing the variance of a Gaussian distribution which 95% C.I. match the observed 95% C.I. width (Fig. 4). G-computation produces the narrowest confidence intervals and as expected their width decreases with increasing sample sizes. The width decrease is more pronounced for DDML. At $n = 250$, DDML produces the widest confidence intervals of all methods whereas for $n = 1,000$, its C.I. width is inferior to the ones obtained with propensity score methods.

Last, we compare results obtained in the scenarios with and without interactions between treatment and covariates. The relative ranking of the different methods remains similar, albeit a difference about DDML relative performance. When $n = 1000$, DDML has a power similar to the propensity-score methods in the linear framework whereas it outperforms these methods in the scenario with an interaction between treatment and covariates (Fig. 4).

Internal replication study

The internal replication study confirms simulation results. G-computation has the smallest MSE and MAE errors for both null and trial replication (Fig. 5, Tables 2 and 3). By contrast the unadjusted approach has the worst performance in terms of MAE and MSE. The two propensity score methods and DDML have intermediate performances (Tables 2 and 3). For null replication, DDML has better performance than IPTW, and PSM has the worst performance (Table 2). For trial replication, DDML has better performance than IPTW, and PSM has the better or worst performance of the three methods depending on the criterion used for evaluation (Table 3).

Width of confidence intervals also varies between methods (Tables 2 and 3). The G-computation method has the smallest width of C.I., the DDML methods has the largest width and the C.I. widths obtained with propensity-score methods are in between. The results mimic what is found at $n = 250$ for the synthetic simulations; the smallest width of C.I is found with G-computation and the largest one is obtained with DDML (Fig. 4).

We also investigate coverage for the null replication (Table 2). The unadjusted method has the lowest coverage (7/9) whereas the propensity-score methods and DDML have complete coverage (9/9). The G-computation has intermediate coverage (8/9) reflecting its narrower confidence intervals.

In terms of estimate and regulatory agreement for the trial replication experiment, DDML has better agreement with trial results followed by G-computation (Table 3). However, differences between the two methods are small; there is regulatory agreement for 16 out of 19 trials with DDML whereas there is regulatory agreement for 15 out of 19 trials with the G-computation method. Agreement with trial results is inferior for propensity-score methods.

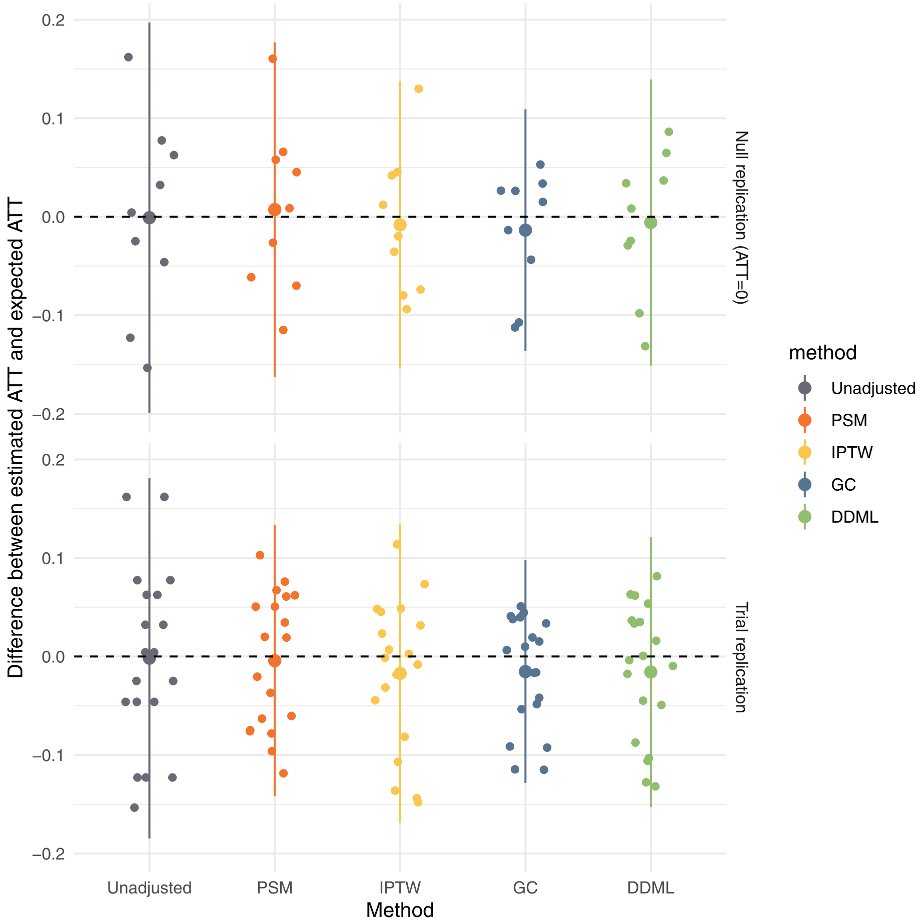

Fig. 5 Results of the two replication experiments. Each point corresponds to an observational experiment. For the 9 null replication experiments, the expected target ATT is 0 and for the 19 RCT experiments, the expected target ATT is the RCT estimate. The larger point is the mean of the points and the bar extends to the mean plus or minus two times the standard deviation. For each method, the position on the x-axis does not matter and random perturbation on the x-axis is added to the points to allow optimal visualisation

Table 2 Results of the negative control experiments when the experimental and control arms are the same. MSE and MAE are respectively the mean squared error and the mean average error between the ATT estimation and the ground truth, which is null. Coverage is the percentage of confidence intervals that contain zero

	MSE(x1000)	MAE(x100)	C.I. width (x1000)	Coverage(%)
Unadjusted	8.73	7.62	24.9	78% (7/9)
PSM	6.46	6.79	28.3	100% (9/9)
IPTW	4.79	5.91	27.9	100% (9/9)
G-computation	**3.53**	**4.79**	22.0	88% (8/9)
DDML	4.72	5.70	28.9	100% (9/9)

Discussion

Based on both synthetic simulations and a replication study of completed randomized trials, we show that statistical methods based on outcome prediction models estimate treatment effect (ATT) more precisely than propensity-score methods, which confirms previous simulation results [19, 21]. Outcome prediction methods have correct type I errors while their power is generally greater than power of propensity score approaches. G-computation methods have increased power compared to propensity score approaches whatever the sample size. The results are more tempered for the DDML approach that explicitly accounts for the bias of machine learning models. For small sample sizes of $n = 250$ individuals, power of DDML can be reduced compared to propensity score methods whereas it is comparable to the power of G-computation for large sample size of $n = 1000$ patients.

Table 3 Results of the RCT replication experiments. Pseudo MSE and MAE are respectively the pseudo mean squared error and the pseudo mean average error obtained by replacing the unknown ground truth with the RCT estimate. Estimate agreement is the percentage of RCT 95% confidence intervals that contain ATT estimation. Regulatory agreement is the percentage of time the cutoff $P < 0.05$ obtained from the non-randomized experiments agrees with the RCT result about $P < 0.05$

	Pseudo MSE(x1000)	Pseudo MAE(x100)	C.I. Width(x100)	Estimate Agreement	Regulatory Agreement
Unadjusted	7.94	7.30	25.1	84.2% (16/19)	73.7% (14/19)
PSM	4.51	6.15	29.0	89.5% (17/19)	73.7% (14/19)
IPTW	5.75	5.86	28.5	89.5% (17/19)	78.9% (15/19)
G-computation	**3.26**	**4.68**	25.9	100% (19/19)	78.9% (15/19)
DDML	4.70	5.60	31.3	100% (19/19)	84.2% (16/19)

There are marked differences between the results obtained with G-computation and DDML. As expected by construction of the DDML estimator, its bias is smaller than the bias of G-computation, which is in turn smaller than the bias of propensity score approaches. Another marked difference concerns the estimation of variance in order to compute confidence intervals. The sample splitting approach overestimates variance of DDML estimator. As a consequence, the widths of confidence intervals for DDML are increased that explains why type I errors are below the 5% threshold rate. DDML being conservative comes at a price of a $10 - 15\%$ reduction of power compared to G-computation when the sample size is small ($n = 250$).

In practice, choosing between DDML and G-computation in a setting with one or several dozens of confounding variable can be guided by sample size. The sample sizes for external control arms can have different orders of magnitude ranging from dozens to thousands of patients [46]. In oncology, after application of inclusion and exclusion criteria, sample size can be smaller than $n = 100$ [47] where G-computation should be preferred, but can also exceed $n = 500$ where DDML can be preferred [48].

A second factor, out of the scope of this paper and potentially influencing the choice between DDML and G-computation, is the dimension of confounding covariates. Our numerical simulations were designed to replicate the current setting of ECA, where adjustment is done on a few clinical and demographic covariates [47]. This scenario with 10-50 covariates and a few hundreds of individuals was also mimicked in our internal replication study. However, in future applications of external control arms, confounding variables can be high dimensional data such as genomics, imaging data, or Electronic Health Record Data [49]. When risk of bias exists because of regularization, the DDML estimator is a promising alternative to G-computation and a comparison of the two estimators in this setting would be insightful.

We trained the propensity score and outcome models with lasso regression and ridge regression respectively. Some studies [50, 51] suggest the potential benefit of using random forests, and boosted CART to estimate the propensity score, reducing bias in treatment effect estimation. The choice of methodology to derive the propensity score and outcome models may impact the relative performance of the methods under comparison and represents an important research direction.

External control arm (ECA) analysis considerably reduces the risks of false positive errors of single arm-trial because it adjusts for the clinical profiles of patients [3]. However, ECA analyses, and more generally RWE analyses, do not fully reproduce results of randomized studies [48, 52, 53]. Therefore, it provides a valuable and less liberal estimation of efficacy than single arm studies [3] but it is not a substitute for large randomized studies. In this paper, we have shown that machine learning methods such as G-Computation and DDML, can improve external control arm analyses by increasing statistical power while preserving type I error.

Conclusions

For external control arm analysis, confounding factors might bias estimation of treatment efficacy because of lack of randomization. To account for confounding factors, propensity score approaches such as IPTW and PSM are the preferred statistical methods. However, our analysis based on synthetic and RCT data shows that methods based on prediction of clinical outcomes, such as G-computation and DDML, are more powerful while having correct type I errors. For a typical ECA setting in oncology with ten confounding covariates and a hundred of patients, G-computation is the most powerful method. More powerful ECA analyses to detect significant treatment effects for newly developed drugs can be obtained by choosing statistical methods based on computational prediction of clinical outcomes.

Table 4 Distribution of main variables used in the replication analysis

Trials	NCT01137812	NCT01106625	NTC01106651	NCT01106677	NCT00968812
Variables					
Age	56.86 (± 9.25)	56.30 (± 9.33)	63.55 (± 6.27)	55.46 (± 9.33)	55.95 (± 9.28)
Sex (Male)	52.3%	6.5%	55%	47%	53%
HBA1C	8.12 (± 0.92)	8.12 (± 0.91)	7.70 (± 0.78)	7.94 (± 0.90)	7.79 (± 0.78)
BMI	33.0 (± 6.49)	31.6 (± 6.76)	31.6 (± 4.60)	31.9 (± 6.10)	31.1 (± 5.35)
LDL	2.70 (± 6.49)	2.54 (± 6.76)	2.33 (± 4.60)	2.78 (± 6.10)	2.66 (± 5.35)
Blood Glucose Level	12.62 (± 7.54)	13.60 (± 6.97)	10.60 (± 2.36)	11.03 (± 13.29)	8.83 (± 1.94)

Appendix

Method to compute DDML estimator and its variance. To take into account the variability of the splitting procedure in the computation of the DDML estimator (Eq. (2)), the split is repeated $S \times K$ times by repeating S K-fold cross-validation procedures. For a cross-validation scheme, the aggregated estimator is the average of the estimators obtained using the k^{th}-fold, $k = 1, \ldots, K$, as the auxiliary dataset

$$\hat{\nu}_s = \frac{1}{K}\sum_{k=1}^{K}\hat{\nu}_{s,k},$$

where $\nu_{s,k}$ is the ATT estimator $A\hat{T}T_{DDML}$ (Eq. (2)) for the s^{th} cross-validation repetition, considering the k^{th} fold as the auxiliary dataset, and the remaining folds to train the two machine learning models. The overall estimate is obtained as

$$\hat{\nu} = \frac{1}{S}\sum_{s=1}^{S}\hat{\nu}_s.$$

The variance of $\hat{\nu}$ is estimated with

$$\hat{\sigma}^2 = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\sigma}_s^2 + \left(\hat{\nu}_s - \hat{\nu}\right)^2\right),$$

where

$$\hat{\sigma}_s^2 = \frac{1}{K}\sum_{k=1}^{K}(\hat{\nu}_{s,k} - \hat{\nu}_s)^2.$$

We always consider 3-fold cross validation. In the Monte Carlo simulations, we consider $S = 20$ repetitions, and in the internal replication, we consider $S = 10$ repetitions.

List of the variable included in the propensity score/outcome model. Serum Albumin, Alkaline phosphatase, Alkaline transaminase, Aspartate transaminase, Basophils/Leukocytes, Biliburin, Blood Urea Nitrogen, Calcium, Cholesterol, Creatine Kinase, Chloride, Serum Creatinine, Eosinophils, Glomerular Filtration Rate Corrected, Gamma-Glutamyl Transferase, Blood sugar level, Plasma Glucose, Hemoglobin A1C, HDL Cholesterol, Hemoglobin, Potassium, LDL, Lymphocytes, Lymphocytes/Leukocytes, Magnesium, Neutrophil, Phosphate, Platelets, Protein, Sodium, Triglycerides, Diastolic Blood Pressure, Systolic Blood pressure, Pulse Rate, Weight, Age, Sex, Race Black or African American, Race other, Race white, Zone asia pacific, Zone central South America, Zone north America, Tabacco use, Concomitant medication diabetes, Previous concomitant medication antihyperglycemic, previous concomitant therapy (Table 4).

Abbreviations

ATT: Average Treatment effect on the Treated; ATE: Average Treatment Effect; C.I. : Confidence Intervals; DDML: Doubly Debiased Machine Learning; ECA: External Control Arm Analysis; IPTW: Inverse Probability of Treatment Weighting; MAE: Mean Absolute Error; MSE: Mean Squared Error; PSM: Propensity Score Matching; RCT: Randomized Clinical Trial; YODA: Yale University Open Data Access.

Acknowledgements

This study, carried out under YODA Project #2019-4077, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C.

Authors' contributions

G.W., M.B., N.L., and P.T. designed the study and experiments. M.H. and M.Z. gave feedback at various stages of the study. N.L. and P.T. performed experiments. M.B. wrote the manuscript that was revised by G.W., M.A., N.L. and P.T. The authors read and approved the final manuscript.

Funding

There is no funding to declare.

Availability of data and materials

Data access should be requested to the Yale University Open Data Access (YODA) Project. Detailed procedure to request data access is provided on YODA website https://yoda.yale.edu/how-request-data.

Declarations

Ethics approval and consent to participate

This manuscript includes a secondary analysis of clinical trial data. The original clinical trials were conducted in accordance with the Declaration of Helsinki. The clinical studies were approved by the relevant review boards, with all patients providing written informed consent. Permission to use the data reported in the secondary analysis of clinical trial data was obtained from the Yale University Open Data Access Project, and subject to a Data Use Agreement. The informed consent of the clinical trial enabled the sharing of data for research that will advance public health.

Consent for publication

Not applicable.

Competing interests

The authors are employees of Owkin, Inc.

Received: 20 January 2022 Accepted: 21 November 2022
Published online: 28 December 2022

References

1. Burcu M, Dreyer NA, Franklin JM, Blum MD, Critchlow CW, Perfetto EM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. Pharmacoepidemiol Drug Saf. 2020;29(10):1228–35.
2. Thorlund K, Dron L, Park JJ, Mills EJ. Synthetic and External Controls in Clinical Trials-A Primer for Researchers. Clin Epidemiol. 2020;12:457.
3. Ventz S, Lai A, Cloughesy TF, Wen PY, Trippa L, Alexander BM. Design and evaluation of an external control arm using prior clinical trials and real-world data. Clin Cancer Res. 2019;25(16):4993–5001.
4. Cassaday RD. When a randomized controlled trial is unlikely: Propensity score analysis of blinatumomab in adults with relapsed/refractory Philadelphia chromosome-positive B-cell acute lymphoblastic leukemia. Cancer. 2020;126(2):253–5.
5. Black N. Why we need observational studies to evaluate the effectiveness of health care. Bmj. 1996;312(7040):1215–8.
6. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet. 2002;359(9302):248–52.
7. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
8. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat. 2002;84(1):151–61.
9. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. Am Stat. 2004;58(4):272–9.
10. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc. 2018;113(521):390–400.
11. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173(7):731–8.
12. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Oxford: Oxford University Press; 2018.
13. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
14. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424.
15. Lenis D, Ackerman B, Stuart EA. Measuring model misspecification: Application to propensity score methods with complex survey data. Comput Stat Data Anal. 2018;128:48–57.
16. Austin PC. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. Multivar Behav Res. 2012;47(1):115–35.
17. Shi C, Blei DM, Veitch V. Adapting Neural Networks for the Estimation of Treatment Effects. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R, editors. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019. Vancouver: NeurIPS. 2019. p. 2503–2513.
18. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol. 2011;173(7):761–7.
19. Chatton A, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Sci Rep. 2020;10(1):1–13.
20. Chatton A, Borgne FL, Leyrat C. Foucher Y. G-computation and doubly robust standardisation for continuous-time data: A comparison with inverse probability weighting. Stat Methods Med Res. 2021;31(4):09622802211047345.
21. McConnell KJ, Lindner S. Estimating treatment effects with machine learning. Health Serv Res. 2019;54(6):1273–82.
22. LaLonde RJ. Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev. 1986;76(4);604–20.
23. Villar PF, Waddington H. Within study comparisons and risk of bias in international development: Systematic review and critical appraisal. Campbell Syst Rev. 2019;15(1–2):e1027.
24. Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? J Econ. 2005;125(1–2):305–53.
25. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014;(4).
26. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92.
27. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012;33(15):1893–901.
28. Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg. 2014;259(1):18–25.
29. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther. 2017;102(6):924–33.
30. Cook TD, Steiner PM. Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. Psychol Methods. 2010;15(1):56.
31. Krumholz HM, Waldstreicher J. The Yale Open Data Access (YODA) project-a mechanism for data sharing. N Engl J Med. 2016;375(5):403–5.
32. Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA Project with clinical trial data sharing after 5 years. Sci Data. 2018;5(1):1–14.
33. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. J Am Stat Assoc. 2005;100(469):322–31.
34. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–60.
35. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Stat Med. 2016;35(30):5642–55.
36. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. Math Model. 1986;7(9–12):1393–512.
37. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput. 1992;4(1):1–58.
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
39. Wilding J, Charpentier G, Hollander P, González-Gálvez G, Mathieu C, Vercruysse F, et al. Efficacy and safety of canagliflozin in patients with type 2 diabetes mellitus inadequately controlled with metformin and sulphonylurea: a randomised trial. Int J Clin Pract. 2013;67(12):1267–82.
40. Schernthaner G, Gross JL, Rosenstock J, Guarisco M, Fu M, Yee J, et al. Canagliflozin compared with sitagliptin for patients with type 2 diabetes who do not have adequate glycemic control with metformin plus sulfonylurea: a 52-week randomized trial. Diabetes Care. 2013;36(9):2508–15.
41. Bode B, Stenlöf K, Sullivan D, Fung A, Usiskin K. Efficacy and safety of canagliflozin treatment in older subjects with type 2 diabetes mellitus: a randomized trial. Hosp Pract. 2013;41(2):72–84.
42. Lavalle-González F, Januszewicz A, Davidson J, Tong C, Qiu R, Canovatchel W, et al. Efficacy and safety of canagliflozin compared with placebo and

sitagliptin in patients with type 2 diabetes on background metformin monotherapy: a randomised trial. Diabetologia. 2013;56(12):2582–92.

43. Cefalu WT, Leiter LA, Yoon KH, Arias P, Niskanen L, Xie J, et al. Efficacy and safety of canagliflozin versus glimepiride in patients with type 2 diabetes inadequately controlled with metformin (CANTATA-SU): 52 week results from a randomised, double-blind, phase 3 non-inferiority trial. Lancet. 2013;382(9896):941–50.
44. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. 2018;47(6):2005–14. https://doi.org/10.1093/ije/dyy120.
45. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized Real-World Evidence to Support Regulatory Decision Making: Process for a Randomized Trial Replication Project. Clin Pharmacol Ther. 2020;107(4):817–26.
46. Goring S, Taylor A, Müller K, Li TJJ, Korol EE, Levy AR, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. BMJ Open. 2019;9(2):e024895.
47. Davies J, Martinec M, Delmar P, Coudert M, Bordogna W, Golding S, et al. Comparative effectiveness from a single-arm trial and real-world data: alectinib versus ceritinib. J Comp Eff Res. 2018;7(09):855–65.
48. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. Clin Pharmacol Ther. 2020;107(2):369–77.
49. Schröder C, Lawrance M, Li C, Lenain C, Mhatre SK, Fakih M, et al. Building external control arms from patient-level electronic health record data to replicate the randomized IMblaze370 control arm in metastatic colorectal cancer. JCO Clin Cancer Inform. 2021;5:450–8.
50. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29(3):337–46.
51. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS ONE. 2011;6(3):e18174.
52. Franklin JM, Glynn RJ, Suissa S, Schneeweiss S. Emulation differences vs. biases when calibrating Real-World Evidence findings against Randomized Controlled Trials. Clin Pharmacol Ther. 2020;107:735–7.
53. Kirchgesner J, Desai RJ, Schneeweiss MC, Beaugerie L, Kim SC, Schneeweiss S. Emulation of a randomized controlled trial in ulcerative colitis with US and French claims data: Infliximab with thiopurines compared to infliximab monotherapy. Pharmacoepidemiol Drug Saf. 2021;31(2):167–75.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.