An (un)Holy Union: Causal Inference, Semiparametric Statistics and Machine Learning in the Age of Data Science

Eric J Tchetgen Tchetgen
Luddy Family President's Distinguished Professor
Professor of Statistics
The Wharton School
University of Pennsylvania

05/09/2023

Outline

- The Role of Causal inference in Epidemiology
- Opportunities for a Holy Union:Causal Inference, Semiparametric Statistical theory and Machine Learning:
    - Selective Machine Learning
    - Credence
    - Conformal Prediction
- Potential Risks for an (Un)Holy Union: The Deconfounder
- Concluding Remarks

Why Causal Inference?

- Central roles of Epidemiology:
  - A Descriptive Role: Distribution of disease frequency;e.g., what is the CVD risk among those who exercise?
  - Causal Inference: Determinants of disease frequency; e.g., does exercise reduce your risk of CVD?
- To estimate the causal effect of an action on an outcome:
  - Physics, chemistry, biology use experimental design
  - Epidemiology, economics, sociology mostly use observational design

Role of Causal Inference in Epidemiology

There are generally two notions of causation in Epidemiology:

- (1)Cause of an effect: first observe an event/outcome, and subsequently identify the causes or events that lead to the observed outcome.
- (2)Effect of a cause: assess the effect of a well defined exposure or intervention. e.g. does smoking cause lung cancer? does AZT prevent the advent of AIDS among HIV infected patients?

Role of Causal Inference in Epidemiology

- An example of (1):In the 80s, when unusual high number of patients dying from a combination of syndromes including a rare Kaposi's skin cancer and pneumonia, the primary scientific objective at the time was to identify the cause of this outbreak. Eventually, HIV found to be the cause.
- Today, I will focus on (2), as most common to biostats and epi methodological research and is relatively easier to address as it does not require complete scientific understanding, although some scientific understanding is certainly needed.
- It formally falls under the experimental paradigm, which is made explicit in the context of randomized experiments, but is still a useful paradigm when experiments cannot be performed for either practical or ethical reasons.e.g. smoking and lung cancer.

Role of Causal Inference in Epidemiology

Why do we need formal theory of causation?

- Makes explicit what we mean by "causal effect", that is what is the quantity/estimand we seek?
- Explains the popular saying " association is not necessarily causation", therefore standard statistical methods may not be used to infer causation.
- Gives conditions under which "association is causation", therefore standard statistical methods may be used.
- Makes explicit assumptions needed for the identification of causal effects, and allows for the derivation of new statistical methods when standard and familiar methods fail.

Role of Causal Inference in Epidemiology

Our causal paradigm consists of:

- Defining causal quantities, typically done in terms of counterfactuals: i.e. What would have been your risk of stroke if, contrary to fact, you had not taken your high blood pressure medication?
- Formulating assumptions sufficient to identify causal quantities (nonparametrically):positivity, consistency and *unconfoundedness*
- Defining a mathematical model to deal with the curse of dimensionality.
- performing statistical inference which includes testing and estimating the magnitude of a causal effect given the observed data.

Why Semiparametric Theory ?

- In well-designed modern observational studies high-dimensional covariate data used to address confounding concerns
  - e.g. Dickerman et al (NEJM, 2022) evaluated comparative effectiveness of mRNA Covid-19 BNT162b2 vs mRNA-1273 in observational study of U.S. Veterans
  - Covariates needed to be adjusted for included:Age, Sex, Race, Ethnicity, Urban residence, Smoking history, long list of comorbidities including lung cancer/CVD/Obesity/Diabetes etc..., Health seeking behavior (PCP visits+# flu vaccines in past 5 yrs)
- Even if willing to assume unconfoundedness given measured factors, how should these factors be accounted for?

Why Semiparametric Theory ?

- Two strategies have emerged over the years.
  - Disease risk modeling strategy: outcome=treatment + risk factors
  - Propensity Score modeling Strategy: treatment= risk factors
- Longstanding debate of which is superior resolved by modern semiparametric theory
  - Do both but carefully$\Longrightarrow$*Double Robustness Protection Principle*: Analyst only needs to do one well but not necessarily both.
  - Also implies that in principle doing both can be substantially better than each separately.

Why Semiparametric Theory?

Double robustness in action: Parametric case

		n=1000		
		PS method	Disease risk method	DR Method
$\alpha_{true}, \eta_{true}$	bias	0.009	0.007	0.009
	variance	0.023	0.022	0.023
	Coverage	0.957	0.950	0.954
$\alpha_{true}, \eta_{false}$	bias	0.009	-0.180	0.01
	variance	0.023	0.020	0.023
	Coverage	0.957	0.747	0.954
$\alpha_{false}, \eta_{true}$	bias	-0.182	0.007	0.009
	variance	0.020	0.022	0.023
	Coverage	0.748	0.950	0.945
$\alpha_{false}, \eta_{false}$	bias	-0.182	-0.180	-0.182
	variance	0.02	0.02	0.02
	Coverage	0.748	0.747	0.741

α:propensity score model

η:Disease Risk model

Why Semiparametric Theory ?

- These simulations confirm that DR theory works in theory and practice, in the sense that it delivers valid causal inference if at least one of two specified strategies can recover valid inference without a priori knowing which model is incorrect.
- However, the last scenario exposes major vulnerability of Double Robustness property in case of parametric models as *all models are wrong in practice*.
- DR paradox:DR estimator may have much larger bias than non-DR estimator in parametric case.
- Acknowledging this paradox invalidates any claim of superior validity made on the basis of double robustness in parametric case.

ML to the rescue ?

- Modern Machine Learning has given rise to several highly flexible methods for constructing predictive models in high dimensional settings; e.g. Random Forests, Gradient Boosting Regression Trees, LASSO, LARS, Elastic Net, Deep Learning etc...
- Disease risk and propensity score are inherently prediction models and therefore prime candidates for flexible ML techniques.
- Major challenge: how to incorporate ML to fit these nuisance models flexibly without compromising statistical guarantees we care about (e.g. Coverage of 95%CI)).
- Two solutions dominate the field:
    - TMLE and its variants (Van der Laan et al)
    - Estimating equations/DDML-type methods (Robins and colleagues, Chernozhukov and colleagues)

ML to the rescue ?

- These ML approaches provide a principled framework for incoporating highly adaptive ML methods to account for high dimensional covariates, however they are tuned to minimize their prediction loss, not necessarily to balance bias and variance of causal effect estimate.
- Specifically, while ML use resolves Double robustness paradox by ensuring that DR estimation bias is dominated by that of non-DR estimators, it does not guarantee good finite sample performance in terms of ensuring bias is as small as it can be given potentially large number of available candidate learners.
- Though CV-TMLE uses stacking allowing for convex combination of learners for optimal performance, this is targeted at reducing prediction error of confounding functions, not necessarily bias of estimated causal effect.

Selective ML for Causal Inference

- To resolve this gap, we have recently proposed an approach for estimating the causal effect of a treatment by adaptively selecting ML adjustment of confounders from a large collection of available machine learners, with the aim of reducing bias.(Cui and TT, 2021)
- Intuitively, Selective ML (Mixed Minimax estimator) settles on the pair of ML estimators of Disease risk and Propensity Score functions that leads to most *stable estimator* of the causal effect in the sense that it is least sensitive to pertubations of each of these functions.

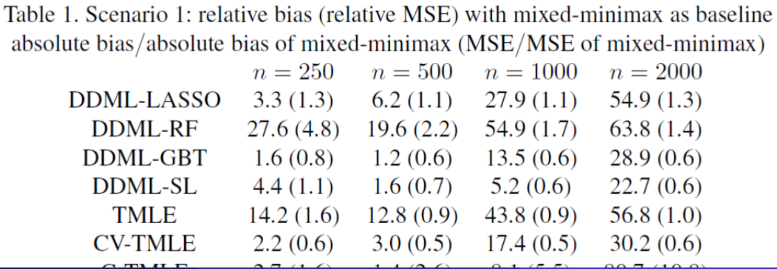


Table 1. Scenario 1: relative bias (relative MSE) with mixed-minimax as baseline absolute bias/absolute bias of mixed-minimax (MSE/MSE of mixed-minimax)

	$n = 250$	$n = 500$	$n = 1000$	$n = 2000$
DDML-LASSO	3.3 (1.3)	6.2 (1.1)	27.9 (1.1)	54.9 (1.3)
DDML-RF	27.6 (4.8)	19.6 (2.2)	54.9 (1.7)	63.8 (1.4)
DDML-GBT	1.6 (0.8)	1.2 (0.6)	13.5 (0.6)	28.9 (0.6)
DDML-SL	4.4 (1.1)	1.6 (0.7)	5.2 (0.6)	22.7 (0.6)
TMLE	14.2 (1.6)	12.8 (0.9)	43.8 (0.9)	56.8 (1.0)
CV-TMLE	2.2 (0.6)	3.0 (0.5)	17.4 (0.5)	30.2 (0.6)
C-TMLE	2.7 (1.6)	1.4 (2.6)	8.1 (5.5)	20.7 (10.8)

Selective ML for Causal Inference

- Selective ML is a promising advance, however does not completely solve the problem at hand: mainly for the data set we wish to analyze, which ML covariate adjustment approach should one select to minimize bias and variance of causal effect estimates? How can results be validated without knowing groundtruth?
- Ideally, would like to perform a *bespoke simulation study* using synthetic data that are stochastically indistinguishable from the observed sample but for which causal groundtruth is available.
- Existing tension between these two potentially conflicting goals as most simulation studies use stylized models of reality in order to know ground truth, while simply resampling from observed sample to simulate data does not provide causal groundtruth against which to evaluate available ML strategies.

A Bespoke Simulation Framework called CREDENCE

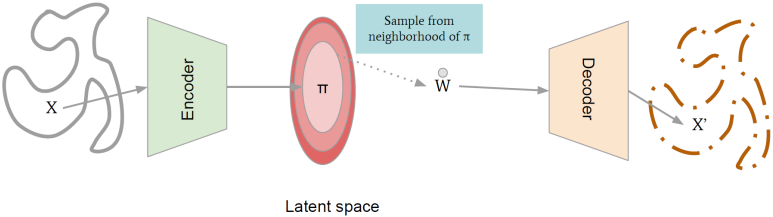
- We have recently introduced a deep generative model-based framework, *Credence*, to validate causal inference methods. (Parikh et al, 2022)
- The framework's novelty stems from its ability to generate synthetic data anchored at the empirical distribution for the observed sample, and therefore virtually indistinguishable from the latter.

A Bespoke Simulation Framework called CREDENCE

- The approach allows the user to specify ground truth for the form and magnitude of causal effects and confounding bias as functions of covariates. Thus simulated data sets are used to evaluate the potential performance of various causal estimation methods when applied to data similar to the observed sample.
- We demonstrated Credence's ability to accurately assess the relative performance of causal estimation techniques in extensive simulation studies and two real-world data applications from Lalonde and Project STAR studies, both have observational and experimental components, making it possible to potentially know groundtruth.

A Bespoke Simulation Framework called CREDENCE

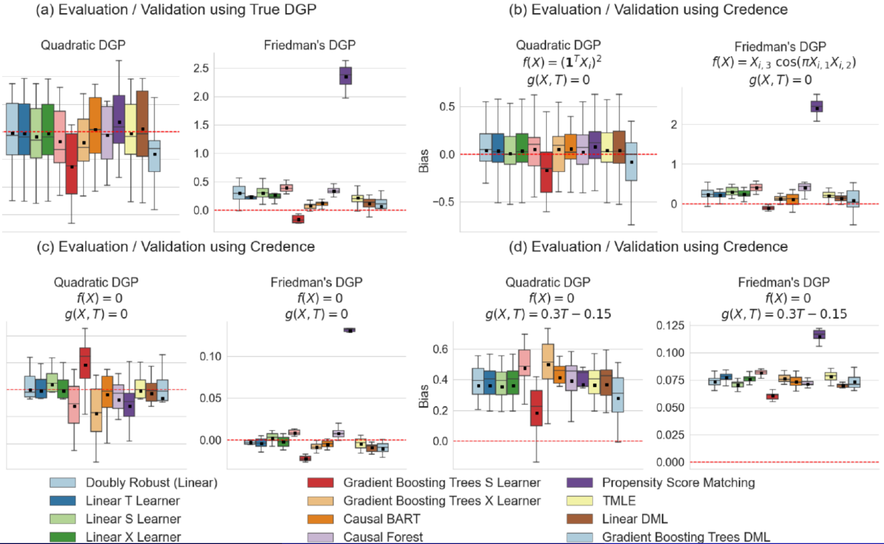
- Credence is a deep variational auto-encoder



with loss function which incorporate standard VAE loss + constraints to shrink causal effect towards user-specified ground-truth f and confounding bias g functions.

$$\mathbf{min}_\theta \left(\begin{array}{c} \mathbf{E}\left[d((X,Y,Z),(X',Y',Z'))\right] \\ +\alpha \left\| \mathbf{E}[Y'(1)-Y'(0)|X'=x'] - f(x') \right\| \\ +\beta \left\| \mathbf{E}[Y'(z')|X'=x',Z'=z'] - \mathbf{E}[Y'(z')|X'=x',Z'=1-z'] - g(x',z') \right\| \end{array} \right)$$

Experimental validation of CREDENCE



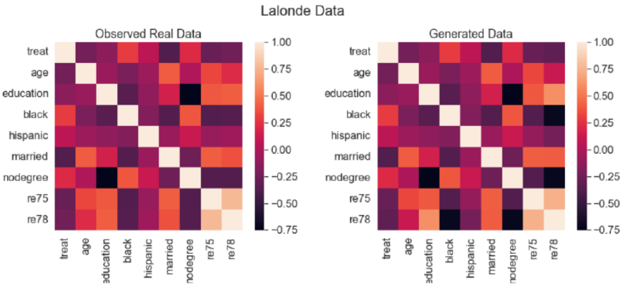
(a) Evaluation / Validation using True DGP

(b) Evaluation / Validation using Credence

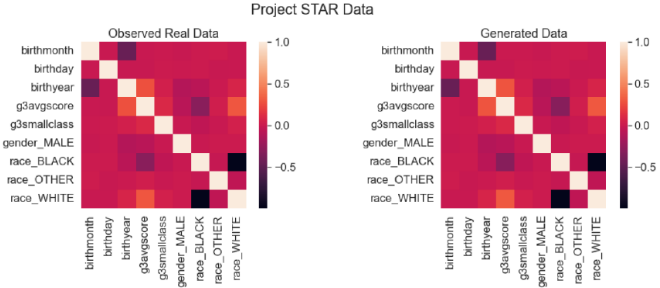
(c) Evaluation / Validation using Credence

(d) Evaluation / Validation using Credence

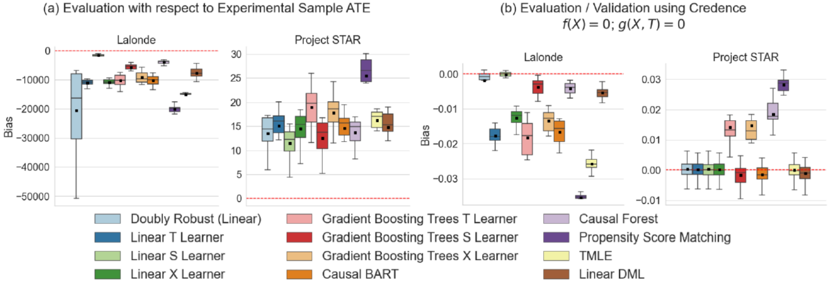
Real World Validation of CREDENCE



(a)Correlation Matrix for real and Credence generated Lalonde data



Real World Validation of CREDENCE



(a) Evaluation with respect to Experimental Sample ATE

(b) Evaluation / Validation using Credence
$f(X) = 0; g(X, T) = 0$

Conformal inference for forecasting individual treatment effects

- Rather than post-hoc validation of various causal inference estimation approaches, ideally would like a method that is guaranteed to provide valid inferences irrespective of whether or not regression models perform well.
- One such recent advance out of machine learning literature is so called *conformal prediction* (Vovk et al, 2005)
- The approach is based solely on an assumption that one has observed N iid samples (X_i, Y_i) and provides an algorithm to construct a well-calibrated prediction interval $\widehat{C}_N$ based on an arbitrary ML prediction algorithm for a new i.i.d test point (X_{N+1}, Y_{N+1}), such that

$$\Pr\left(Y_{N+1} \in \widehat{C}_N\left(X_{N+1}\right)\right) \geq 1-\alpha$$

- What is remarkable is that the guaranty is finite sample exact, irrespective of the quality of the ML prediction algorithm. Of course, the better the forecast algorithm, the tigher the resulting set

Conformal inference for forecasting individual treatment effects

- Rather than post-hoc validation of various causal inference estimation approaches, ideally would like a method that is guaranteed to provide valid inferences irrespective of whether or not regression models perform well.
- One such recent advance out of machine learning literature is so called *conformal prediction* (Vovk et al, 2005)

Conformal inference for forecasting individual treatment effects

- The approach is based solely on an assumption that one has observed N iid samples (X_i, Y_i) and provides an algorithm to construct a well-calibrated prediction interval $\widehat{C}_N$ based on an arbitrary ML prediction algorithm for a new i.i.d test point (X_{N+1}, Y_{N+1}), such that
$$\Pr\left(Y_{N+1} \in \widehat{C}_N\left(X_{N+1}\right)\right) \geq 1-\alpha$$
- What is remarkable is that the guaranty is finite sample exact, irrespective of the quality of the ML prediction algorithm. Of course, the better the forecast algorithm, the tigher the resulting set $\widehat{C}_N\left(\left(X_{N+1}\right)\right)$.

Conformal inference for forecasting individual treatment effects

- Conformal inference recently extended to handle so-called *covariate shift setting*, whereby the target sample for which we seek predictions has different covariate distribution than training sample for which outcome labels are available (Tibshirani et al, 2019).
- The approach requires knowing the likelihood ratio of covariate densities for each sample and using it as a weight in weighted conformal prediction algorithm.
- The original weighted conformal prediction (WCP) does not account for uncertainty about estimated weights.

Conformal inference for forecasting individual treatment effects

- Causal inference formulation of this problem is to generate well-calibrated prediction interval for so-called individual treatment effect (ITE), i.e. such that in large samples

$$\Pr\left(Y_{N+1}(1) - Y_{N+1}(0) \in \widehat{C}_N\left(X_{N+1}\right)\right) \geq 1 - \alpha$$

- Lei and Candes (2021) recently established that WCP has asymptotic control of coverage probability accounting for estimation of likelihood ratio. Notably, the error rate of their coverage is the minimum of two sources of bias, that of the forecast learning algorithm used to construct conformal score, and the other of the covariate density ratio.

Conformal inference for forecasting individual treatment effects

- We have recently developed a generic doubly robust prediction interval for ITE which presents several advantages over existing methods (Yang et al, 2022, Qiu et al, 2022).
- First, the error of coverage probability is guaranteed to shrink at a rate faster than that established by Lei and Candes (2021) in nonparametric context;i.e. the error rate is a product of estimation errors as opposed to minimum.

Conformal inference for forecasting individual treatment effects

- Also, our approach is generic in the sense of allowing arbitrary learners for adjusting for confounding/treatment switching and for estimating a conformal score.
- Finally we adopt the approach of Yang and Kuchibhotla (2021) to select among different training algorithms and identifies a prediction region with simultaneous valid coverage and (approximately) optimal width formally expressed in terms of an oracle inequality .

Conformal inference for forecasting individual treatment effects (Yang et al 2022)

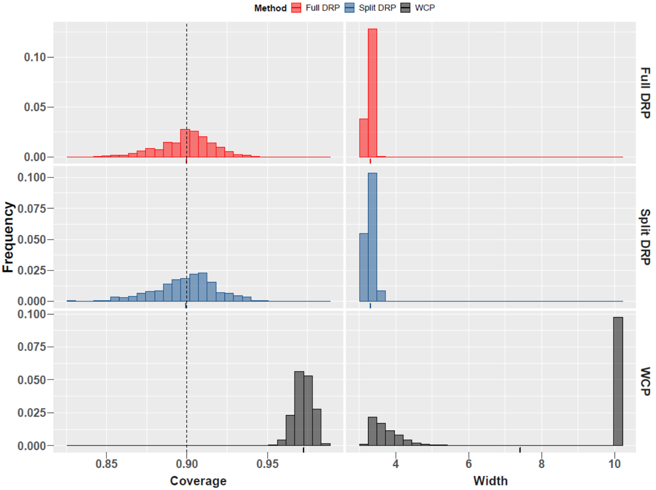


Figure 2: Histograms of coverage and width of Doubly Robust Prediction (DRP) and Weighted Conformal Prediction (WCP) on synthetic data using the absolute residual score. The width is truncated at 10 for

Conformal inference for forecasting (Qiu et al, 2022)

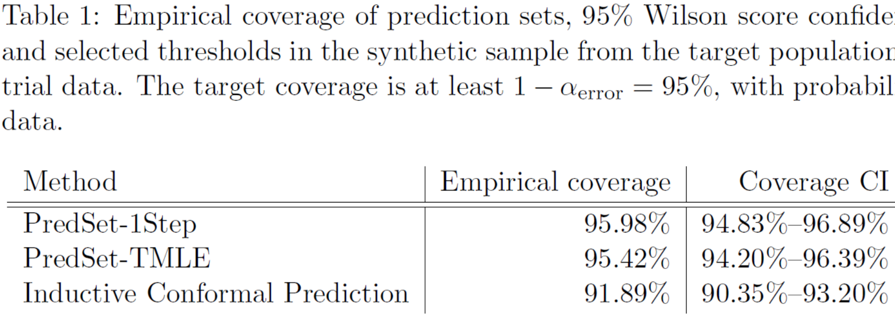


Table 1: Empirical coverage of prediction sets, 95% Wilson score confide and selected thresholds in the synthetic sample from the target populatio trial data. The target coverage is at least $1 - \alpha_{\text{error}} = 95\%$, with probabil data.

Method	Empirical coverage	Coverage CI
PredSet-1Step	95.98%	94.83%–96.89%
PredSet-TMLE	95.42%	94.20%–96.39%
Inductive Conformal Prediction	91.89%	90.35%–93.20%

Potential Risks for an (Un)Holy Union: The Deconfounder

- I have discussed three case studies at the nexus of interactions between causal inference, semiparametric statistics and Machine Learning
- Guiding principles of these case studies is to ground identification and inference on well-established principles of causal inference and statistical inference, why safely leveraging flexibility, adaptive ability and richness of ML tools.
- Note however that causal inference is hard! Ultimately relies upon untestable assumptions that cannot be ruled out or confirmed strictly empirical basis.Specifically, thus far we have assumed that there is no hidden bias by unmeasured confounders.

Potential Risks for an (Un)Holy Union: The Deconfounder

- Beware of methods that claim identification without explicitely relying on such untestable assumption.
- E.g. The deconfounder of Wang and Blei (2019) that claims in the multiple treatment setting, to learn about and correct for any residual confounding by repurposing the joint treatments to construct a confounding control variate using ML tools.
- In other words, all of the relevant information about selection into treatment that pertains to confounding, can be completely recovered by the realized treatment.

Potential Risks for an (Un)Holy Union: The Deconfounder

- To be true, such claims must ultimately rely on hidden assumptions or a degenerate model of state of nature, otherwise, cannot belong to what Pearl refers to as the second rung in the ladder of causation (BOW, Pearl and Mackenzie, 2018);
- see Ogburn et al (2019, 2020), D'Amour (2019a,b) and Grimmer et al (2020) for a deconstruction of the Deconfounder claims; The last paper shows that naive OLS generally outperforms deconfounder!
- Also see Miao et al (2016), TT et al (2021) for recent development on a principled approach for *proximal causal inference* leveraging proxies of unmeasured confounders.

Acknowledgments

- This work is funded by NIH grants R01AG065276,R01CA222147 and R01AI27271. (PI:Tchetgen Tchetgen)
- My collaborators : Yifan Cui (NUS), Harsh Parikh (Duke), Louise Xu (Amazon), Carlos Varjao (Amazon), Yachong Yang (Wharton), Arun Kuchibhotla (CMU)