Federated learning for Causal Inference using deep generative disentangled models

Alejandro Almodóvar, Juan Parras, Santiago Zazo
Information Processing and Telecommunication Center, Universidad Politécnica de Madrid
alejandro.almodovar@upm.es

UNIVERSIDAD POLITÉCNICA DE MADRID
POLITÉCNICA
ETSIT UPM
NEURAL INFORMATION PROCESSING SYSTEMS

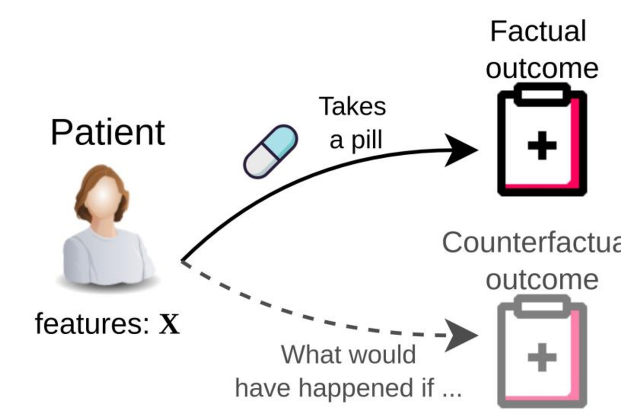
Task: Counterfactual prediction

Classical causal inference: Estimate individual treatment effects (ITEs) of a treatment (drug or medication) over a target variable (T → Y).

For binary outcomes:

$$ITE_i = Y_i(T_i = 1) - Y_i(T_i = 0)$$

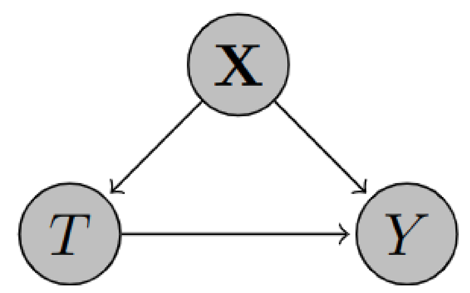
- Factual outcome: observed.
- Counterfactual has to be predicted.



Methodology for ITE estimation

- Block the backdoor path (through X)
- Estimate conditional average treatment effect: CATE

$$\hat{ITE}_i = \mathbb{E}[Y|T=1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y|T=0, \mathbf{X} = \mathbf{x}_i]$$



Arising challenge

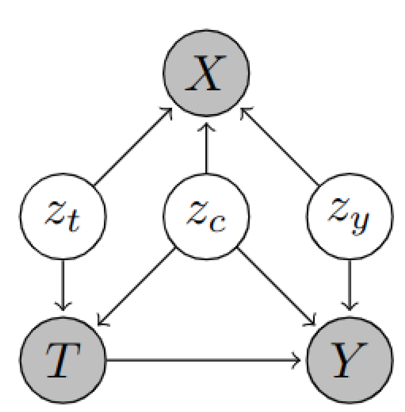
- In real data is difficult to know the causal graph
- How we verify that CI assumptions are met?

Causal Inference standard assumptions
1) Unconfoundedness
2) Positivity
3) No interference
4) Consistency

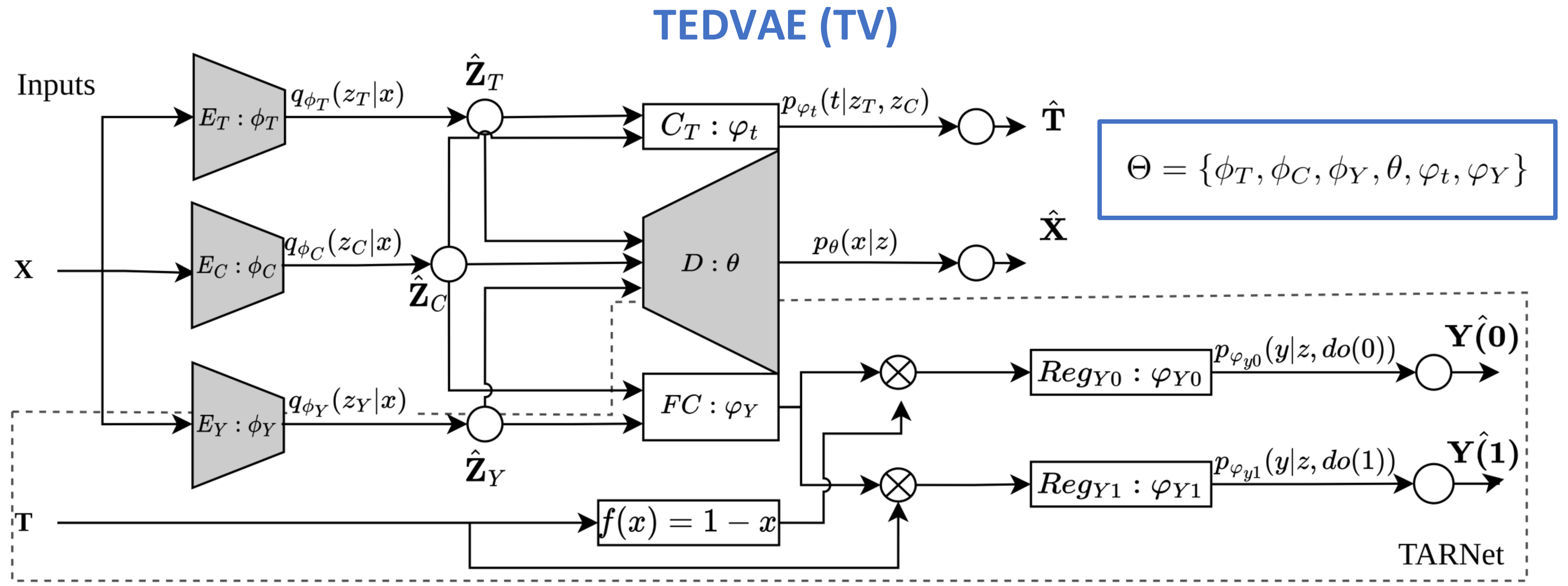
Local Causal Inference model: TEDVAE

Treatment Effect Distentangled Variational Autoencoder [2]

- Achieves a partial discovery of the causal graph
- Latent factors divide the covariates contributions into:
  - Instrumental variables, Z_t
  - Confounders, Z_c
  - Adjustment variables, Z_y
- Predict causal effects only from confounders and adjustment variables reduces the bias and the variance



TEDVAE (TV)



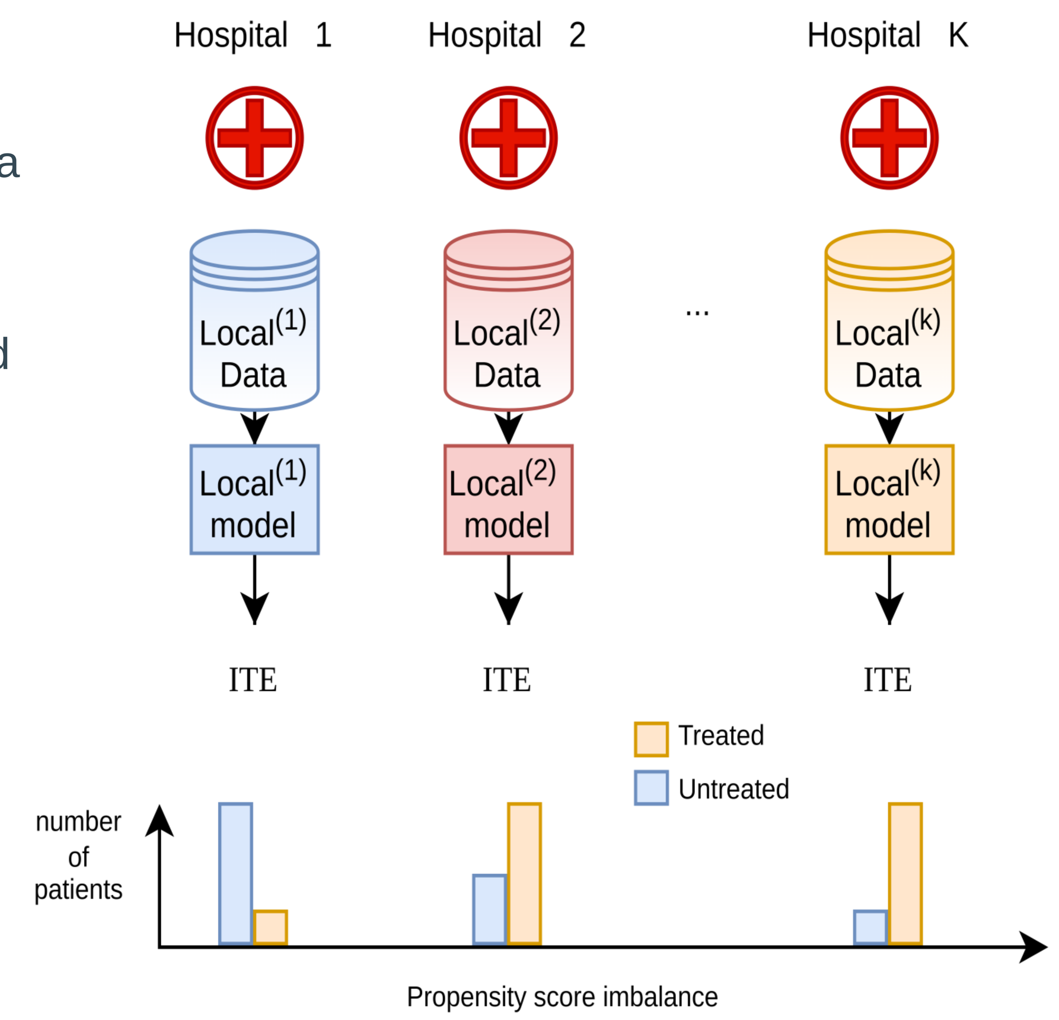
$$L_{\mathrm{TEDVAE}}(\Omega; \mathcal{D}) = \frac{1}{N} \sum_{i \in \mathcal{M}} l_{\mathrm{ELBO}}(\mathbf{x}_i, y_i, t_i; \Theta) + \alpha_t \mathbb{E}_{q_{\phi_t} q_{\phi_c}} [\log p_{\varphi_t}(t_i \mid \mathbf{z}_{t,i}, \mathbf{z}_{c,i})] + \alpha_y \mathbb{E}_{q_{\phi_y} q_{\phi_c}} [\log p_{\varphi_y}(y_i \mid t_i, \mathbf{z}_{c,i}, \mathbf{z}_{y,i})]$$

$$\begin{aligned} l_{\mathrm{ELBO}}(\mathbf{x}, y, t; \Theta) = & \mathbb{E}_{q_{\phi_c} q_{\phi_t} q_{\phi_y}} [\log p_\theta(\mathbf{x} \mid \mathbf{z}_t, \mathbf{z}_c, \mathbf{z}_y)] \\ & - D_{KL}(q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) \| p_{\theta_t}(\mathbf{z}_t)) \\ & - D_{KL}(q_{\phi_c}(\mathbf{z}_c \mid \mathbf{x}) \| p_{\theta_c}(\mathbf{z}_c)) \\ & - D_{KL}(q_{\phi_y}(\mathbf{z}_y \mid \mathbf{x}) \| p_{\theta_y}(\mathbf{z}_y)) \end{aligned}$$

Distributed causal inference

Decentralized data

- Several hospitals (nodes)
- Each hospital has its own data
- Each hospital has a CI model
- Privacy constraints
- Patient data cannot be shared



New distributed CI conditions

1) Same set of covariates
2) Stable covariate distribution

$$p^j(\mathbf{X}) = p^k(\mathbf{X})$$

3) Stable propensity score

$$p^j(T|\mathbf{X}) = p^k(T|\mathbf{X})$$

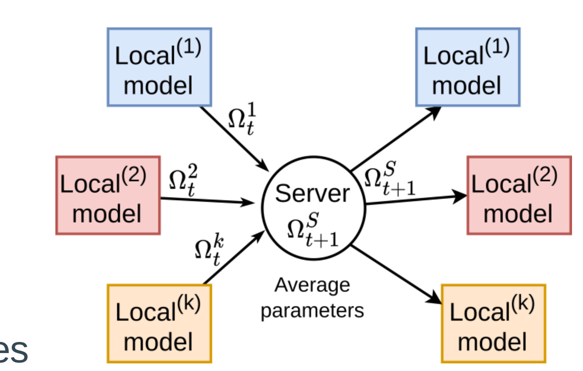
Our problem

- Conditions 2 and 3 do not hold
- Some hospitals have a more restricted access to some medication

Propensity score adaptation of `FedAvg`

Federated Averaging: `FedAvg`

- Share model parameters, not patients' information
- Iterate until convergence:
  1) Train several epoch each node independently
  2) Share local model parameters to a central server
  3) Average the parameters, weighting by the number of samples



$$\Omega^S_{t+1} = \sum_{k=1}^{K} \frac{N^k}{N} \Omega^k_{t+1}$$

Propensity adaptation (PA)

- Outcome regressors parameters are weighted by the number of treated/control patients in each node.

Vanilla `FedAvg`	*Propensity Adaptation* of `FedAvg`
$\Omega^S_{t+1} = \left\{ \begin{array}{l} \sum_{k=1}^K \frac{N^k}{N} \Theta^k_{t+1}, \\ \sum_{k=1}^K \frac{N^k}{N} \varphi^k_{Y1_{t+1}}, \\ \sum_{k=1}^K \frac{N^k}{N} \varphi^k_{Y1_{t+1}} \end{array} \right\}$	$\Omega^S_{t+1} = \left\{ \begin{array}{l} \sum_{k=1}^K \frac{N^k}{N} \Theta^k_{t+1}, \\ \sum_{k=1}^K \frac{N^k_T}{N^S_T} \varphi^k_{Y1_{t+1}}, \\ \sum_{k=1}^K \frac{N^k_C}{N^S_C} \varphi^k_{Y1_{t+1}} \end{array} \right\}$

Evaluation

Semisynthetic benchmarking data: IHDP datasets

- Characteristics → Length: 747 samples, X: 25 covariates
- Setting A: Both potential outcomes are linear combination of $\mathbf{X}$ and T
- Setting B: Y_0 is linear combination and Y_1 is exponential combination of $\mathbf{X}$ and T

Evaluation metric: Precision Error of Heterogeneous Effects

$$\sqrt{\mathrm{PEHE}} = \sqrt{\mathbb{E}\left[(\hat{\mathrm{ITE}}(x) - \mathrm{ITE}(x))^2\right]}$$

Results

Baselines

- TV centralized: all data is combined in a single node (violates privacy constraints)
- TV isolated: each node trains in isolation with its own local data
- TV FedAvg Vanilla (V): standard implementation of FedAvg, without PA
- FedCI: Federated Causal inference from [3]
- CausalRFF: Causal Random Fourier Features from [4]

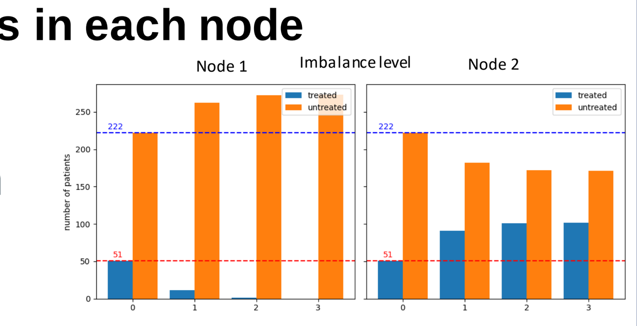
Small sample of the original distribution

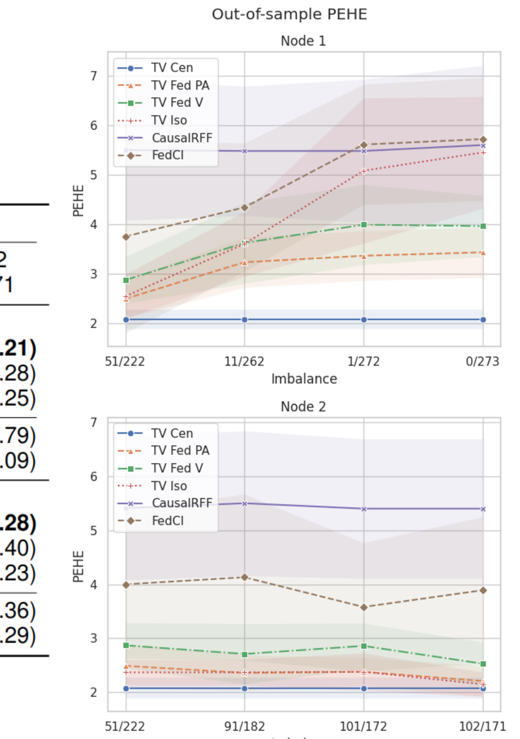
- 2 nodes, 83 patients in each node
- Same propensity score in each node
- *Vanilla* and PA `FedAvg` works similar
- TEDVAE outperforms other methods

	Setting A		Setting B	
	node 1	node 2	node 1	node 2
TV Cen	1.16(0.26)		3.07(0.72)	
TV Fed PA	**1.18(0.31)**	**1.20(0.31)**	**3.55(0.86)**	**3.41(0.69)**
TV Fed V	1.15(0.37)	1.15(0.29)	3.61(0.80)	3.50(0.72)
TV Iso	1.21(0.41)	1.27(0.29)	4.83(0.81)	4.64(0.65)
CausalRFF	2.99(1.73)	2.96(1.72)	6.88(1.39)	6.80(1.37)
FedCI	2.56(0.45)	2.63(0.83)	4.88(1.95)	4.94(2.16)

Increasing the imbalance of treated/control patients in each node

- 2 nodes, 273 patients in each node
- start from the original distribution of the data at each node
- Progressive increase in propensity score imbalance
- PA FedAvg is the one that works best after Centralized TV in very unbalanced cases.





		Imbalance 0		Imbalance 1		Imbalance 2		Imbalance 3	
		Node 1 51/222	Node 2 51/222	Node 1 11/262	Node 2 91/182	Node 1 1/272	Node 2 101/172	Node 1 0/273	Node 2 102/171
Setting A	TV Cen	0.75(0.17)							
	TV Fed PA	**0.74(0.25)**	**0.79(0.23)**	**0.68(0.15)**	**0.78(0.25)**	**0.86(0.25)**	**0.85(0.26)**	**0.90(0.30)**	**0.81(0.21)**
	TV Fed V	0.73(0.25)	0.78(0.25)	0.75(0.15)	0.84(0.22)	1.04(0.30)	0.94(0.22)	1.21(0.41)	0.95(0.28)
	TV Iso	0.85(0.25)	0.82(0.27)	0.81(0.23)	0.83(0.18)	2.56(0.86)	0.81(0.19)	2.89(1.02)	0.85(0.25)
	CausalRFF	2.93(1.48)	2.82(1.52)	2.94(1.49)	2.88(1.79)	3.25(1.79)	2.95(1.89)	3.35(1.92)	3.25(1.79)
	FedCI	1.89(1.00)	2.36(1.32)	2.13(0.98)	1.68(0.98)	3.35(1.66)	1.75(0.89)	3.55(1.82)	2.85(2.09)
Setting B	TV Cen	2.08(0.19)							
	TV Fed PA	**2.49(0.39)**	**2.47(0.41)**	**3.23(0.51)**	**2.36(0.33)**	**3.36(0.50)**	**2.38(0.26)**	**3.43(0.51)**	**2.21(0.28)**
	TV Fed V	2.87(0.47)	2.86(0.42)	3.62(0.82)	2.71(0.56)	3.99(0.81)	2.86(0.42)	3.96(0.62)	2.53(0.40)
	TV Iso	2.54(0.45)	2.37(0.24)	3.60(0.66)	2.37(0.24)	5.08(1.47)	2.38(0.35)	5.45(1.13)	2.15(0.23)
	FedCI	3.75(1.95)	4.00(1.37)	4.34(1.30)	4.13(1.54)	5.61(1.22)	3.58(1.19)	5.72(1.25)	3.89(1.36)
	CausalRFF	5.50(1.42)	5.41(1.33)	5.48(1.31)	5.50(1.34)	5.48(1.45)	5.40(1.29)	5.60(1.61)	5.40(1.29)

Out-of-sample PEHE results with increasing imbalances

Conclusion & Next steps

- Propensity adaptation metrics are between Centralized TV and the other methods
- Test the algorithm with more nodes and other imbalances.
- Test the algorithm with more replications of IHDP and other datasets: ACIC, TWINS and synthetic data.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017), "Communication-efficient learning of deep networks from decentralized data," in Artificial intelligence and statistics, pp. 1273–1282, PMLR.
[2] W. Zhang, L. Liu, and J. Li (2020), "Treatment effect estimation with disentangled latent factors," in AAAI Conference on Artificial Intelligence.
[3] T. V. Vo, Y. Lee, T. N. Hoang, and T.-Y. Leong (2022), "Bayesian federated estimation of causal effects from observational data," in Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (J. Cussens and K. Zhang, eds.), vol. 180 Proceedings of Machine Learning Research, pp. 2024–2034, PMLR, 01–05.
[4] T. V. Vo, A. Bhattacharyya, Y. Lee, and T.-Y. Leong (2022), "An adaptive kernel approach to federated
learning of heterogeneous causal effects," Advances in Neural Information Processing Systems, vol. 35, pp. 24459–24473.

Paper PDF