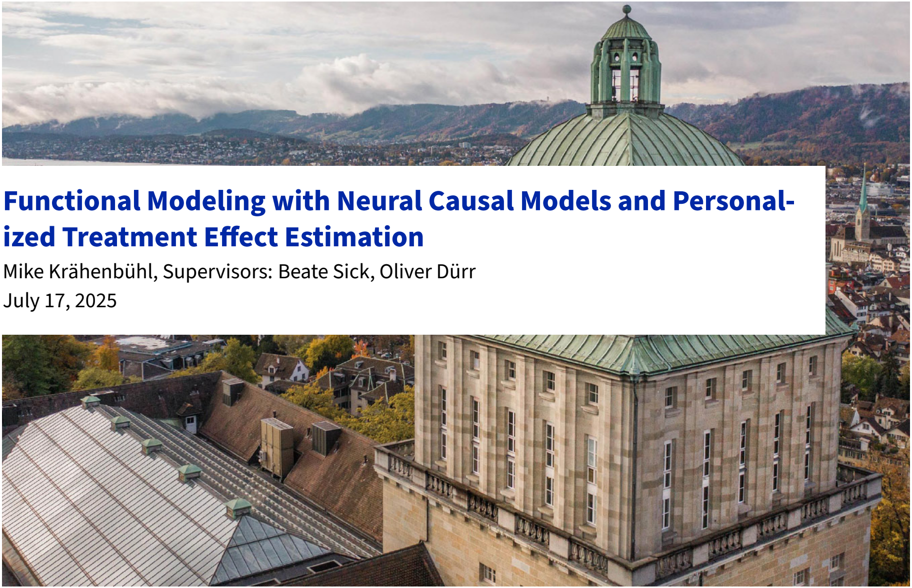


Functional Modeling with Neural Causal Models and Personalized Treatment Effect Estimation

Mike Krähenbühl, Supervisors: Beate Sick, Oliver Dürr

July 17, 2025

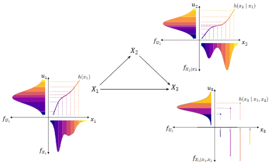
Background

Supervisors:

- Beate Sick, UZH
- Oliver Dürr, HTWG Konstanz

Paper *"Interpretable Neural Causal Models with TRAM-DAGs"* (Sick and Dürr, 2025):

- Framework to model causal relationships
- Based on transformation models
- Rely on (deep) neural networks
- Compromise between interpretability and flexibility



They showed on synthetic data, that TRAM-DAGs can be fitted on observational data and tackle causal queries on all three levels of Pearl's causal hierarchy.

Research Questions

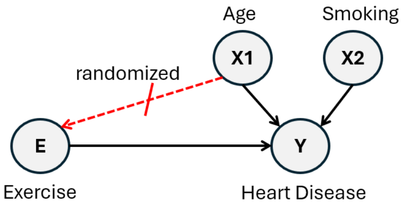
In this presentation:

1. TRAM-DAGs
   - How do they work?
2. Individualized Treatment Effect (ITE) estimation
   - Does it work on real data (International Stroke Trial)?
   - When and why does ITE estimation fail (simulation)?
   - How to estimate ITEs with TRAM-DAGs in a complicated graph (simulation)?

RCT vs. Observational Data

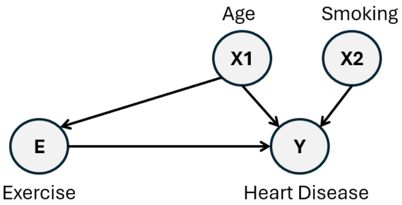
Randomized Controlled Trial:

- Gold standard for estimating causal effect
- Solves problem of confounding



Observational Data:

- Real world, potential confounding
- We assume no unobserved confounding

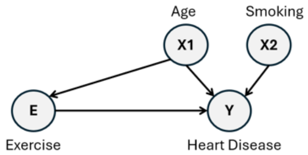


Pearl's Causality Ladder

Observational (seeing)

$P(Y = 1 \mid E = 1)$

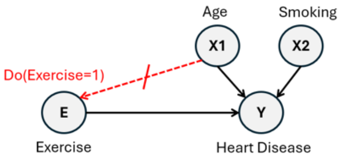
"Probability of heart disease given that the person exercises"



Interventional (doing)

$P(Y = 1 \mid \mathrm{do}(E = 1))$

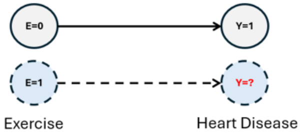
"Probability of heart disease if we made people start exercising"



Counterfactual (imagining)

$P(Y_{(E=1)} = 1 \mid E = 0, Y = 1)$

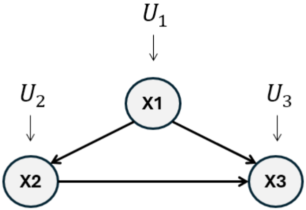
"Would someone who does not exercise and has heart disease still have it if they had exercised?"



Structural Causal Model

SCM: Describes the causal mechanism and probabilistic uncertainty

- X_i = observed variable
- U_i = noise distribution



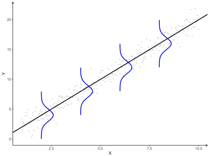
$$U_1 \sim F_{U_1}, U_2 \sim F_{U_2}, U_3 \sim F_{U_3}$$

$$X_1 = f_1(U_1)$$
$$X_2 = f_2(U_2, X_1)$$
$$X_3 = f_3(U_3, X_1, X_2)$$

Estimating Functional Form

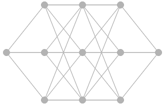
Statistical methods:

- E.g. linear/logistic regression
- Predefined form, risk of bias if misspecified



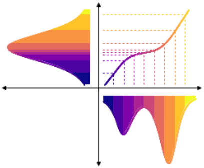
Neural networks:

- E.g. feed-forward NNs, normalizing flows, VACAs
- Flexible, but "black-box", data-type limitations



TRAM-DAGs:

- Compromise: flexibility + interpretability
- Mixed data types



Transformation Models

Flexible distributional regression method (Hothorn et al., 2014)

Continuous $Y \in \mathbb{R}$:

$$F_{Y|\mathbf{X}=\mathbf{x}}(y) = F_Z(h(y) + \mathbf{x}^\top \beta)$$

Discrete $Y \in \{y_1, y_2, \ldots, y_K\}$:

$$P(Y \leq y_k \mid \mathbf{X} = \mathbf{x}) = F_Z(\vartheta_k + \mathbf{x}^\top \beta), \quad k = 1, 2, \ldots, K-1$$

- F_Z: CDF of the standard logistic distribution
- h: Transformation function, monotonically increasing
- $\mathbf{x}$: Predictors

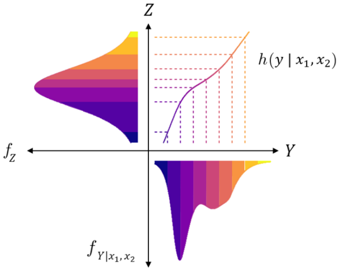
Transformation Models

Continuous Y:

Intercept: Bernstein polynomial

$h_I(y) = \frac{1}{M+1} \sum_{k=0}^{M} \vartheta_k \, \mathrm{B}_{k,M}(y)$

$h(y \mid \mathbf{x}) = h_I(y) - \mathbf{x}^\top \boldsymbol{\beta}$

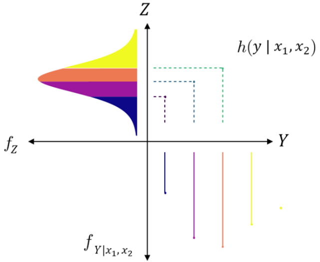


Discrete/Ordinal Y:

Intercept: Cut-off value

$h_I(y_k) = \vartheta_k$

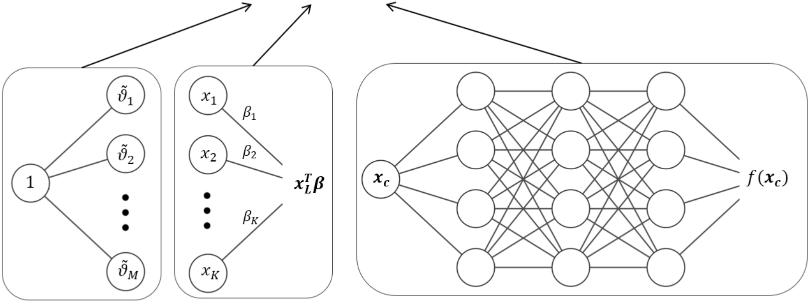
$h(y_k \mid \mathbf{x}) = h_I(y_k) - \mathbf{x}^\top \boldsymbol{\beta}$



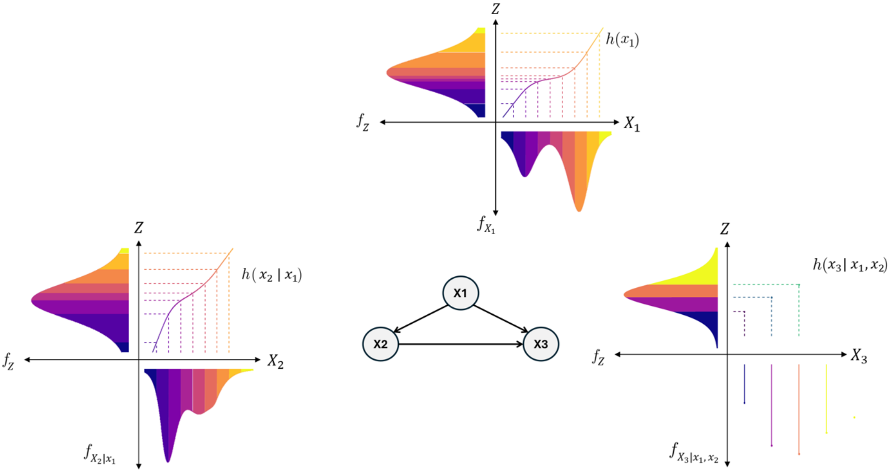
Deep TRAMs

- Extended to Deep TRAMs (Sick et al., 2021)
- Flexible components
- Minimize the NLL through NN optimization

$$h(y|\boldsymbol{x_L}, \boldsymbol{x_C}) = \mathrm{SI} + \mathrm{LS}_{\boldsymbol{x_L}} + \mathrm{CS}_{\boldsymbol{x_C}}$$



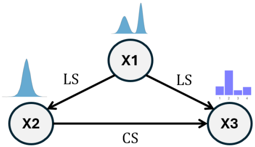
TRAM-DAGs



Z
$h(x_1)$
f_Z
X_1
f_{X_1}
Z
$h(x_2 \mid x_1)$
f_Z
X_2
$f_{X_2|x_1}$
X1
X2
X3
Z
$h(x_3 \mid x_1, x_2)$
f_Z
X_3
$f_{X_3|x_1,x_2}$

Simulation Example

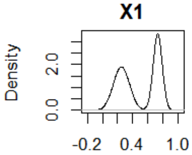
- We have:
  - Observational data (simulated)
  - Predefined DAG
- We want:
  - Estimate conditional CDF of each variable
  - Sample from conditional distributions for causal queries with structural equations $x_i = h^{-1}(z_i \mid \mathrm{pa}(x_i))$



$$X_1 \sim F_z(h(x_1))$$
$$X_2 \sim F_z(h(x_2) + \mathrm{LS}_{x1})$$
$$X_3 \sim F_z(h(x_3) + \mathrm{LS}_{x1} + \mathrm{CS}_{x2})$$

Data Generating Process (DGP)

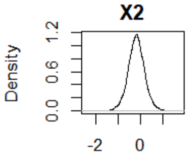
X_1**:** Continuous, bimodal. *Source node* (independent).



X_2**:** Continuous. Depends on X_1 (linear):

$$\beta_{12} = 2, \quad h_I(X_2) = 5X_2$$

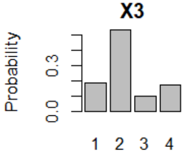
$$\boxed{h(X_2 \mid X_1) = h_I(X_2) + \beta_{12}X_1}$$



X_3**:** Ordinal. Depends on X_1 (linear) and X_2 (complex):

$$\beta_{13} = 0.2, \quad f(X_2) = 0.5 \cdot \exp(X_2), \quad \vartheta_k \in \{-2,\ 0.42,\ 1.02\}$$

$$\boxed{h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13}X_1 + f(X_2)}$$



Construct Model: Modular Neural Network

Inputs:

Observations + assumed structure

Outputs:

- Simple Intercepts (SI): ϑ
- Linear Shifts (LS): $\beta_{12}X_1$, $\beta_{13}X_2$
- Complex Shift (CS): $f(X_2)$

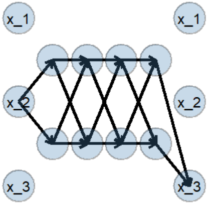
Transformation Functions:

$$\boxed{h(X_i \mid pa(X_i)) = \mathrm{SI} + \mathrm{LS} + \mathrm{CS}}$$

$$h(X_1) = h_I(X_1)$$

$$h(X_2 \mid X_1) = h_I(X_2) + \beta_{12}X_1$$

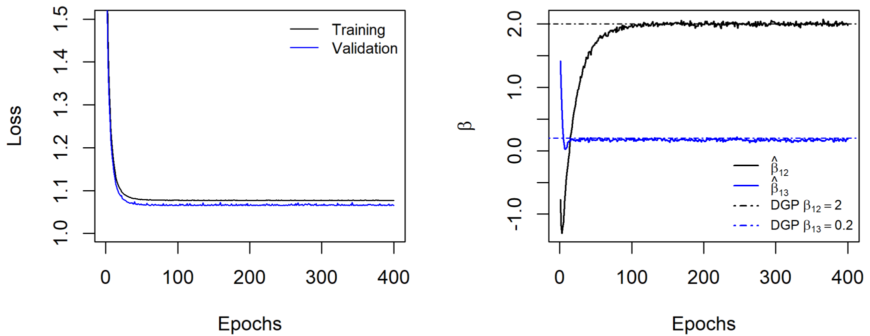
$$h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13}X_1 + f(X_2)$$



CS_{X_2} on X_3

Experiment 1: TRAM-DAGs (model learning)

20,000 training samples



Sampling from the Fitted TRAM-DAG (observational)

Nodes $X_i, i \in \{1, 2, 3\}$**:**

- Sample latent value:

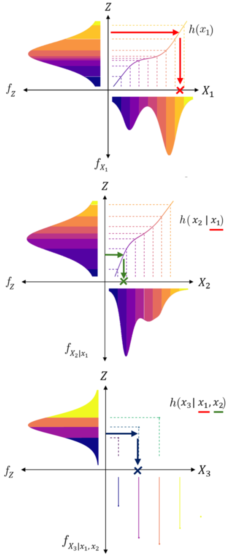
$$z_i \sim F_{Z_i} \quad \text{(e.g., \texttt{rlogis()} in R)}$$

- Determine x_i such that:
  - **If X_i is continuous:** Solve for x_i using numerical root-finding:

$$h(x_i \mid \mathrm{pa}(x_i)) - z_i = 0$$

  - **If X_i is ordinal:** find the smallest category x_i such that

$$x_i = \max\left(\{0\} \cup \{x : z_i > h(x \mid \mathrm{pa}(x_i))\}\right) + 1$$

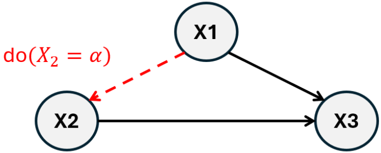


Sampling from the Fitted TRAM-DAG (interventional)

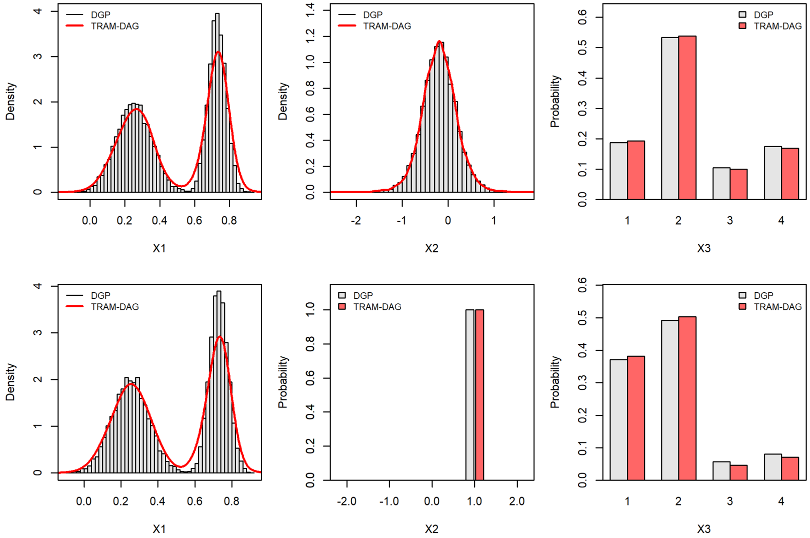
Interventional sampling:

- Do-intervention: $\mathrm{do}(x_2 = \alpha)$
- Sample from the interventional-distribution:

$$x_3 = \min\{x : z_3 \leq h(x \mid x_1, x_2 = \alpha)\}$$

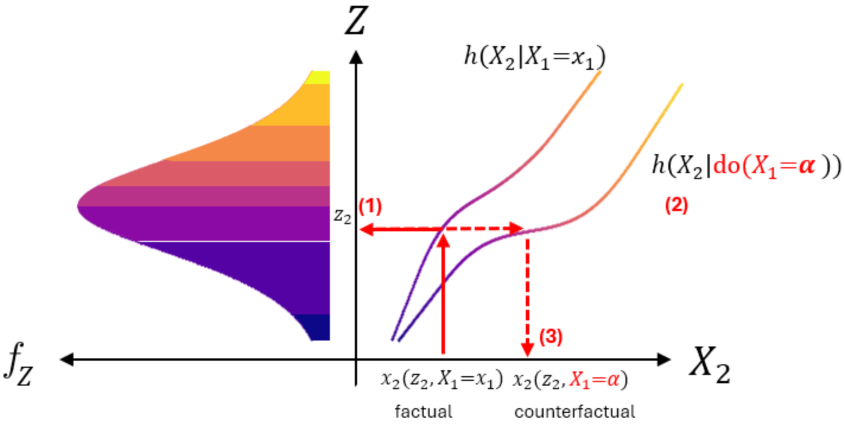


Experiment 1: TRAM-DAGs (sampling distributions)



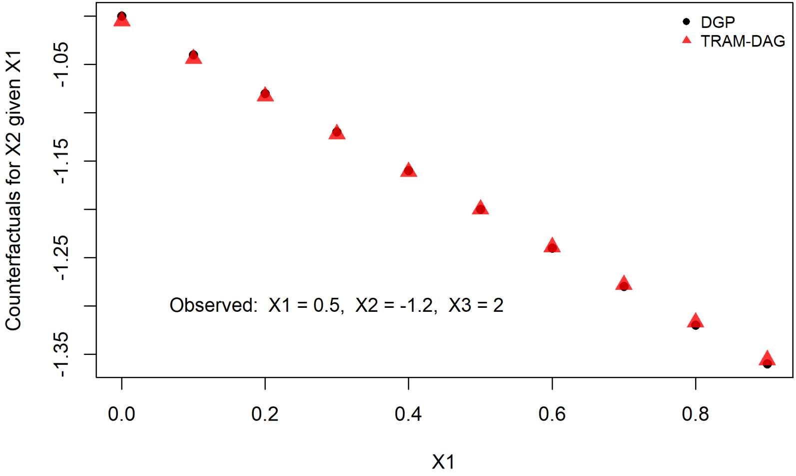
Experiment 1: TRAM-DAGs (counterfactuals)

How to determine a counterfactual value for X_2, given some observation?



Experiment 1: TRAM-DAGs (counterfactuals)

Counterfactuals: Counterfactual value of X_2 under varying X_1



Experiment 1: TRAM-DAGs (Discussion)

With TRAM-DAGs we can:

- Estimate the functional form of the edges in the DAG
- Customize flexibility (SI/CI, LS, CS)
- Sample from the fitted model
- Estimate counterfactuals

Individualized Treatment Effect (ITE): Motivation

Motivation:

— RCT typically estimates Average Treatment Effect (ATE)
— Individuals may respond differently depending on characteristics
— Crucial for decision-making in personalized medicine or targeted marketing
— Heterogeneous treatent effect mainly due to treatment-covariate-interactions

Individual treatment effect: Difference in potential outcomes

$$Y_i(1) - Y_i(0)$$

, where $Y_i(1)$ is the potential outcome if treated and $Y_i(0)$ if not treated.

Fundamental problem of causal inference $\rightarrow$ We cannot observe both potential outcomes for the same individual.

Individualized Treatment Effect (ITE): Assumptions

Assumptions for identifiability of causal effects from observed data:

1. **Consistency:** Observed outcome equals the potential outcome under the treatment actually received: $Y = Y(1)$ if $T = 1$, and $Y = Y(0)$ if $T = 0$
2. **Ignorability/Unconfoundedness:** Treatment assignment is independent of potential outcomes given observed covariates: $(Y(1), Y(0)) \perp T \mid X$
3. **Overlap/Positivity:** Every individual has a positive probability of receiving each treatment level: $0 < P(T = 1 \mid X = x) < 1 \quad$ for all x.
4. **No interference:** The treatment of one individual does not affect the potential outcomes of another individual.

Individualized Treatment Effect (ITE): Estimand

If assumptions for identifiability are satisfied:

$$\begin{aligned}
\mathrm{ITE}_i(\mathbf{x}_i) &= \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}_i] \\
&= \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}_i] \\
&= \mathbb{E}[Y_i(1) \mid T_i = 1, \mathbf{X}_i = \mathbf{x}_i] - \mathbb{E}[Y_i(0) \mid T_i = 0, \mathbf{X}_i = \mathbf{x}_i] \quad \text{(by ignorabilit} \\
&= \mathbb{E}[Y_i \mid T_i = 1, \mathbf{X}_i = \mathbf{x}_i] - \mathbb{E}[Y_i \mid T_i = 0, \mathbf{X}_i = \mathbf{x}_i] \quad \text{(by consistency)}
\end{aligned} \tag{1}$$

For a binary outcome:

$$\mathrm{ITE}_i(\mathbf{x}_i) = P(Y_i = 1 \mid T_i = 1, \mathbf{X}_i = \mathbf{x}_i) - P(Y_i = 1 \mid T_i = 0, \mathbf{X}_i = \mathbf{x}_i). \tag{2}$$

Individualized Treatment Effect (ITE): Models

How we estimated the potential outcomes?

- T-learners: Two separate models, estimated on treated and control groups (logistic regression, tuned random forest)
- S-learners: One model, with treatment as a feature (TRAM-DAGs)

Experiment 2: ITE on International Stroke Trial (IST)

Chen et al. (2025) showed that results of models used for ITE estimation did not generalize to the test set.

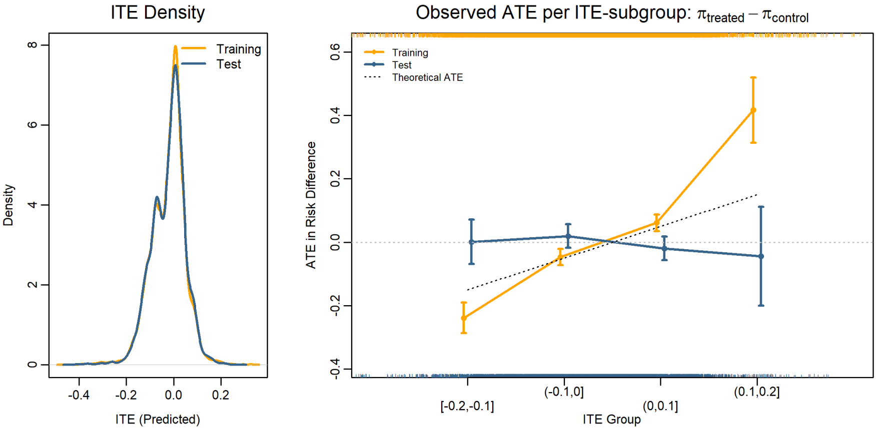
International Stroke Trial (IST):

— Large RCT on stroke patients (19,435 patients, 21 baseline covariates)
— Evaluated the effects of aspirin on stroke patients
— Binary treatment and outcome

Goal: Estimate ITE with T-learners (logistic regression, tuned random forest) and S-learner (TRAM-DAGs) on IST data.

Experiment 2: ITE on International Stroke Trial (IST): Results

Results with T-learner tuned random forest (comets package):



Experiment 2: ITE on International Stroke Trial (IST): Discussion

Our interpretation:

- Similar results as Chen et al. (2025)
- Some models suggest a range of ITEs, but do not generalize to the test set (no effect)

Experiment 3: Simulation of ITE estimation robustness

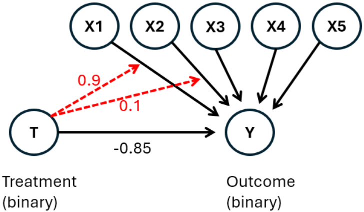
Goal:

- Simulate different RCT scenarios to understand when ITE estimation fails
- Apply simple model (logistic regression) and complex model(tuned random forest)

Simulation Case 1: Fully Observed

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction

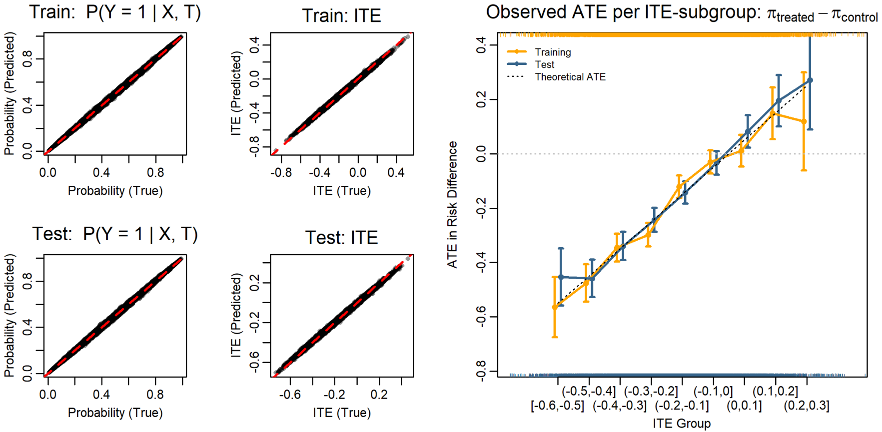


Outcome model:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

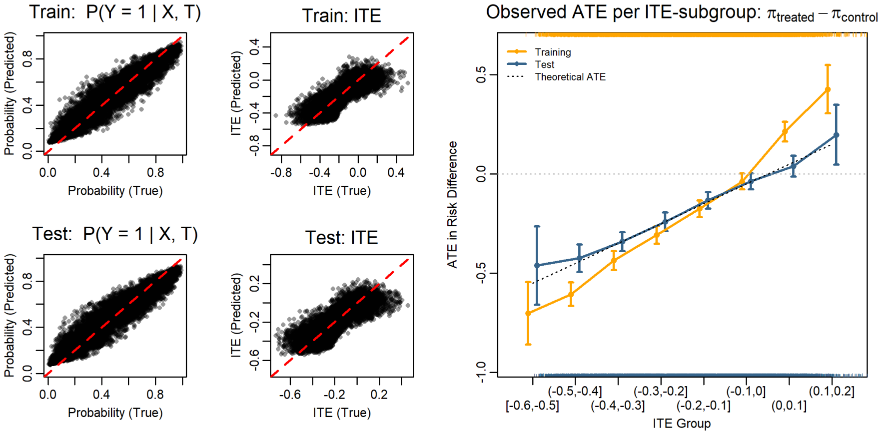
Simulation Case 1: Fully Observed

Results with T-learner logistic regression (glm):



Simulation Case 1: Fully Observed

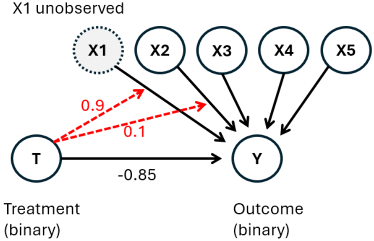
Results with T-learner Random Forest (comets package):



Simulation Case 2: Unobserved Interaction

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction



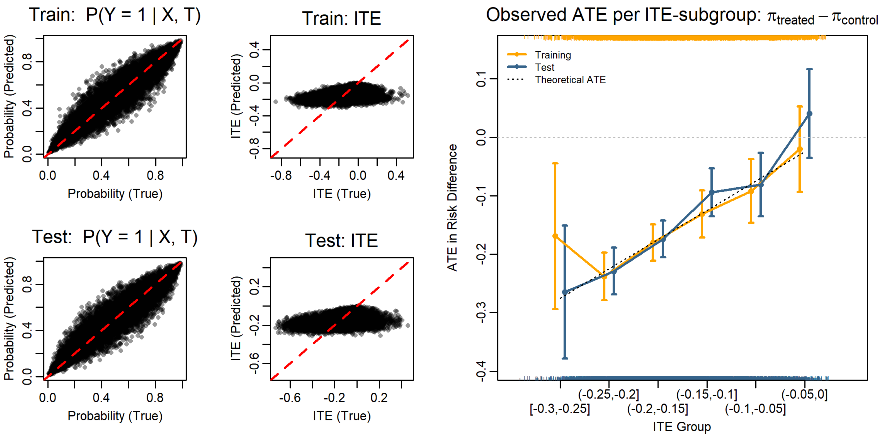
Outcome model:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

Note: Same DGP, but X_1 is not observed!

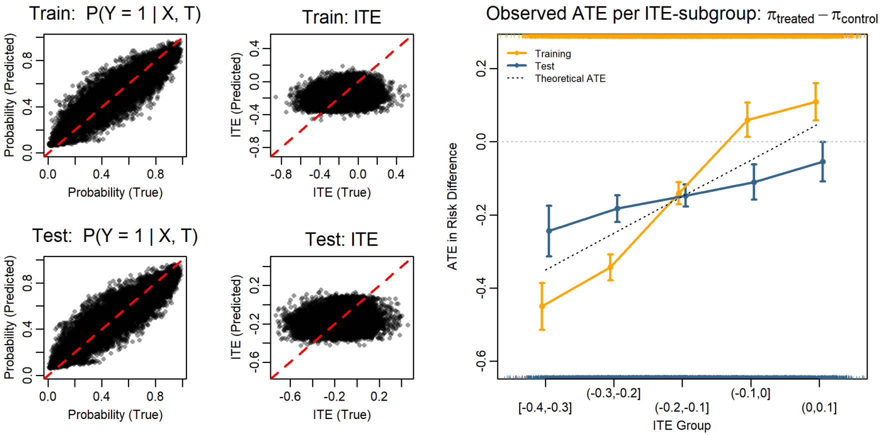
Simulation Case 2: Unobserved Interaction

Results with T-learner logistic regression (glm):



Simulation Case 2: Unobserved Interaction

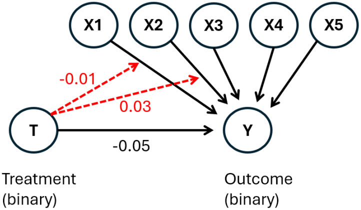
Results with T-learner Random Forest (comets):



Simulation Case 3: Fully Observed, Small Effects

Setup:

- $n = 20{,}000$
- $T \sim \text{Bernoulli}(0.5)$
- $\mathbf{X} = (X_1, \ldots, X_5)^\top \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- $\mathbf{X_{TX}} = (X_1, X_2)^\top$ interaction



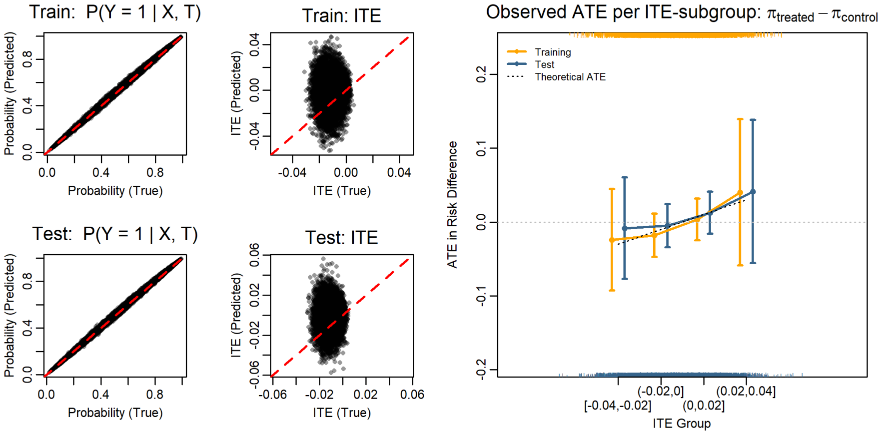
Outcome model:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}, T) = \text{logit}^{-1}\left(\beta_0 + \beta_T T + \boldsymbol{\beta}_X^\top \mathbf{X} + T \cdot \boldsymbol{\beta}_{TX}^\top \mathbf{X_{TX}}\right)$$

Note: Same DGP, but weak treatment effects!

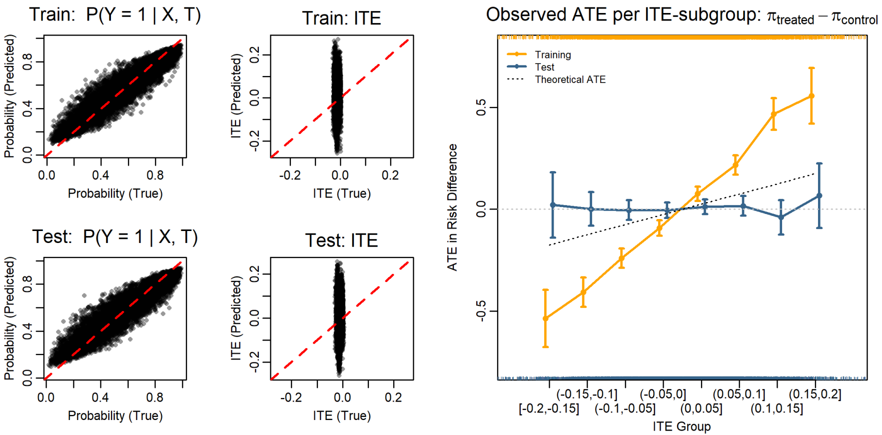
Simulation Case 3: Fully Observed, Small Effects

Results with T-learner logistic regression (glm):



Simulation Case 3: Fully Observed, Small Effects

Results with T-learner Random Forest (comets package):

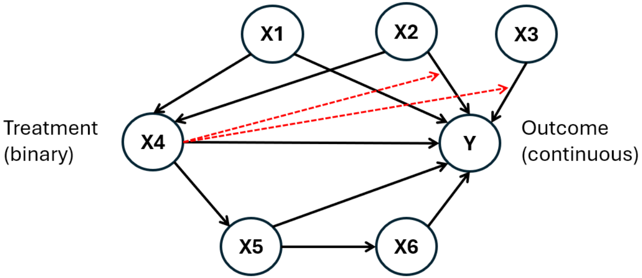


ITE simulation: Interpretation

My interpretation:

- When a high predicted treatment effect (ITE) corresponds to a high observed effect in the train set (strong discrimination), but not in the test set, it might be due to **unobserved interaction variables** or **weak treatment effects**.
- This is more likely to occur with complex models, as they tend to overfit when the interaction is not observed.

TRAM-DAGs: Example for ITE estimation

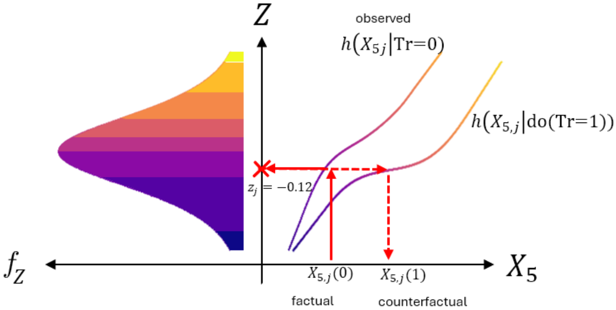


DGP:

- $X5 = h_5^{-1}(\epsilon - 0.8\,X4)$ $\rightarrow$ (depends on treatment)
- $X6 = h_6^{-1}(\epsilon + 0.5\,X5)$ $\rightarrow$ (depends on treatment through X5)
- $Y = h_7^{-1}(\epsilon - \beta_1 X1 - \beta_2 X2 - \beta_3 X3 - \beta_4 X4 - \beta_5 X5 - \beta_6 X6 - Tr \cdot (\beta_{2Tr} X2 + \beta_{3Tr} X3))$

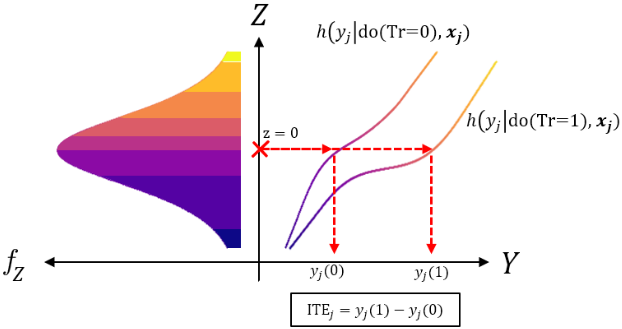
TRAM-DAGs: Estimate Potential Outcomes

If we observe a $X5$ under $Tr = 0$, we can determine the counterfactual $X5$ under $Tr = 1$ with the observed latent value z_j:



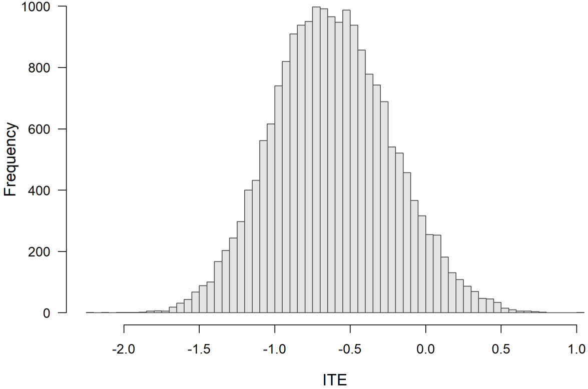
TRAM-DAGs: Example for ITE estimation

$$\text{ITE} = \text{median}(Y \mid \text{do}(T = 1), X) - \text{median}(Y \mid \text{do}(T = 0), X)$$



TRAM-DAGs: Example for ITE estimation

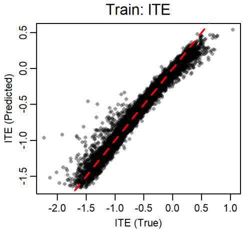
$$\text{ITE} = \text{median}(Y \mid \text{do}(T=1), X) - \text{median}(Y \mid \text{do}(T=0), X)$$

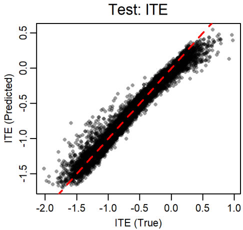


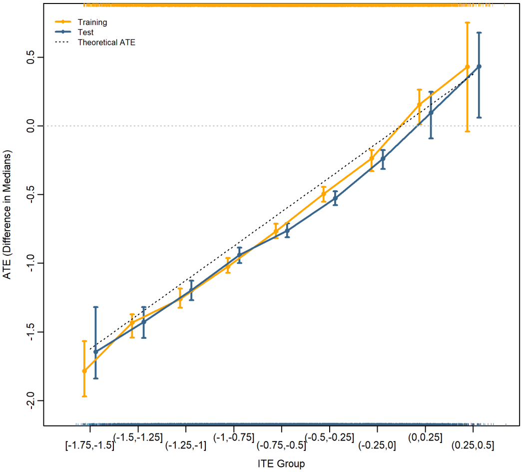
TRAM-DAGs: Estimate Potential Outcomes

1. Estimate each $h_i(X_i \mid \mathrm{pa}(X_i))$ fully flexible (deep-NN / complex intercept)
2. Take the train set or a test set
3. $Z_i = h(X_i \mid pa(X_i))$ gives us the (observed) latent variable for each X_i
4. Determine counterfactuals for X5 and X6 with the (observed) latent variables Z_i
5. Determine medians of potential outcomes $Y(1)$ and $Y(0)$
6. $\mathrm{ITE} = \mathrm{median}(Y(1) \mid X_{tx}) - \mathrm{median}(Y(0) \mid X_{ct})$

TRAM-DAGs: Example for ITE estimation (Results)







TRAM-DAGs: Example for ITE estimation (Results)

ATE TRAM-DAG: estimated as mean($\text{ITE}_{predicted}$):

-0.619 (-0.627 to -0.617)

ATE from RCT (randomized:) estimated as
observed median($Y \mid T = 1$) - median($Y \mid T = 0$):

-0.637 (-0.662 to -0.610)

- confidence intervals obtained by bootstrapping

References

Chen, H., Aebersold, H., Puhan, M. A., and Serra-Burriel, M. (2025). Causal machine learning methods for estimating personalised treatment effects – insights on validity from two large trials.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):3–27.

Sick, B. and Dürr, O. (2025). Interpretable neural causal models with tram-dags. Accepted at the CLeaR 2025 Conference.

Sick, B., Hathorn, T., and Dürr, O. (2021). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2476–2481.