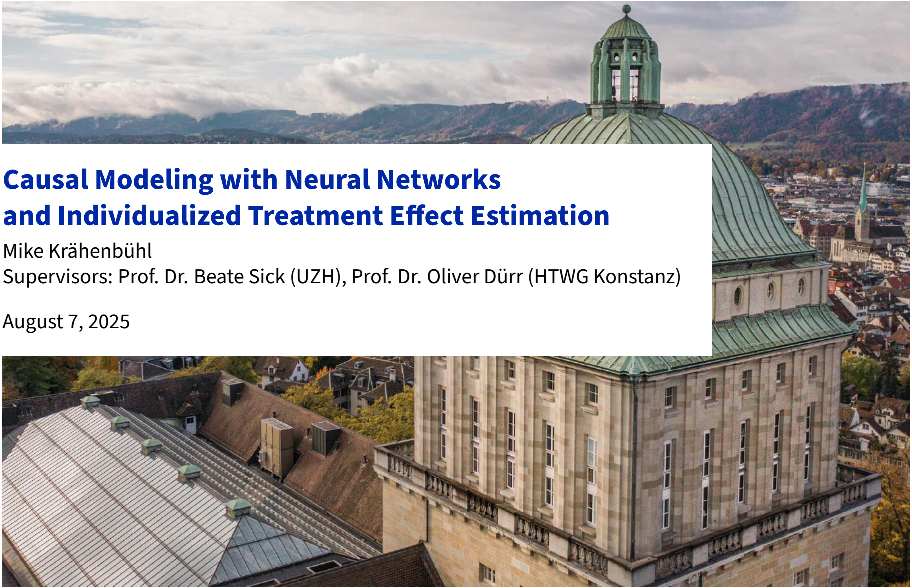
Universität Zürich[UZH]

Master Program in Biostatistics www.biostat.uzh.ch
Master Exam



Causal Modeling with Neural Networks and Individualized Treatment Effect Estimation

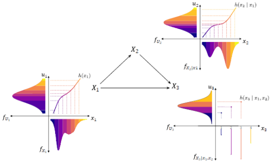
Mike Krähenbühl
Supervisors: Prof. Dr. Beate Sick (UZH), Prof. Dr. Oliver Dürr (HTWG Konstanz)

August 7, 2025

Background

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

- Framework to model causal relationships in a known directed acyclic graph (DAG)
- 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 to fit the model on observed data and subsequently make observational, interventional and counterfactual queries?
2. Individualized Treatment Effect (ITE) estimation
   - Does ITE estimation work on real RCT data (International Stroke Trial)?
   - When and why does ITE estimation fail (simulation)?
   - How to estimate ITEs with TRAM-DAGs in a complicated graph (simulation)?

TRAM-DAGs

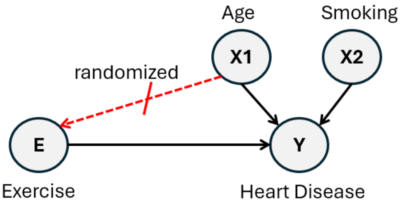
TRAM-DAGs: Motivation

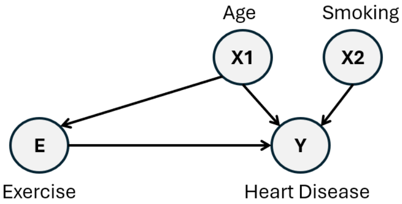
Randomized Controlled Trial:

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

Observational Data:

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



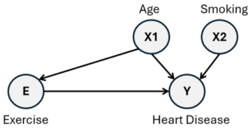


TRAM-DAGs: Motivation

Pearl's causal hierarchy (Pearl, 2009)

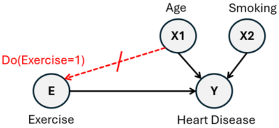
(L1) Observational: $P(Y = 1 \mid E = 1)$

"Probability of heart disease given that the person exercises"



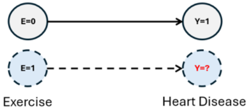
(L2) Interventional: $P(Y = 1 \mid \mathrm{do}(E = 1))$

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



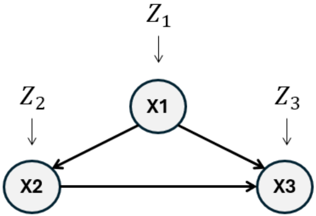
(L3) Counterfactual: $P(Y_{(E=1)} = 1 \mid E = 0, Y = 1)$

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



TRAM-DAGs: Background

Structural Causal Model: Describes the causal mechanism and probabilistic uncertainty (Pearl, 2009)



$$Z \sim F_{Z_1}, Z_2 \sim F_{Z_2}, Z_3 \sim F_{Z_3}$$

$$X_1 = f_1(Z_1)$$

$$X_2 = f_2(Z_2, X_1)$$

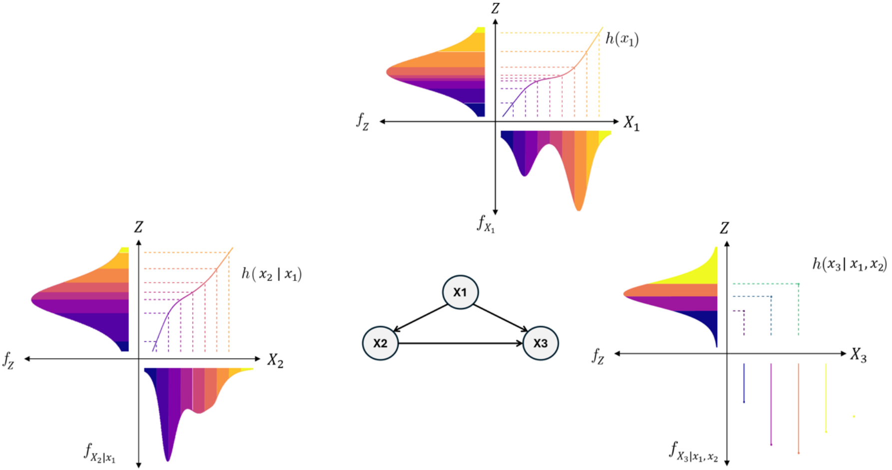
$$X_3 = f_3(Z_3, X_1, X_2)$$

- X_i : observed variable
- Z_i : exogeneous (latent) variable
- f_i : deterministic function: $X_i = f_i(Z_i, \text{pa}(X_i))$

$\rightarrow$ We want a model that estimates $X_i = f_i(Z_i, \text{pa}(X_i))$ in a flexible and interpretable way!

TRAM-DAGs: Background

Proposed framework: TRAM-DAGs (Sick and Dürr, 2025)



TRAM-DAGs: Background

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

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

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

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 latent distribution (e.g. standard logistic)
- h: Transformation function, monotonically increasing
- $\mathbf{x}$: Predictors

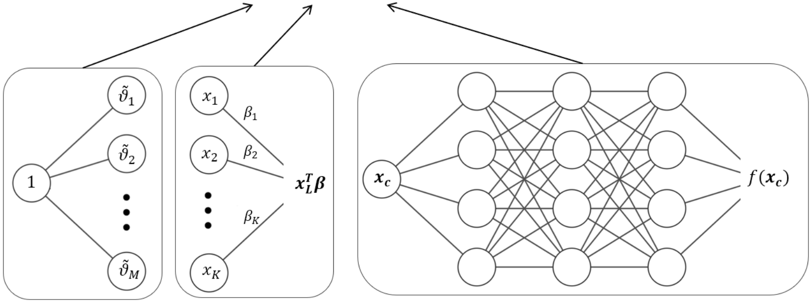
TRAM-DAGs: Background

Extended to Deep TRAMs (Sick et al., 2021)

- Customizable transformation model using neural networks (NNs)
- Minimizing negative log-likelihood (NLL) via NN optimization

Effects of predictors: LS (Linear Shift), CS (Complex Shift), CI (Complex Intercept)

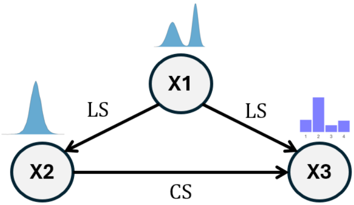
$$h(y|\boldsymbol{x}_L, \boldsymbol{x}_C) = \mathrm{SI} + \mathrm{LS}_{\boldsymbol{x}_L} + \mathrm{CS}_{\boldsymbol{x}_C}$$



TRAM-DAGs: Experiment 1 (Simulation)

Setup:

- Observational data (simulated)
- Predefined DAG



$$h(X_1) = h_I(X_1)$$
$$h(X_2 \mid X_1) = h_I(X_2) + \beta_{12} X_1$$
$$h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13} X_1 + f(X_2)$$

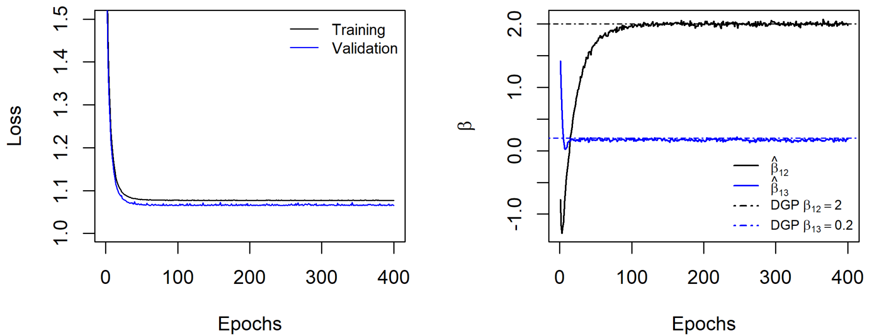
$$\boxed{f(X_2) = 0.5 \cdot \exp(X_2)}$$

We want:

- With TRAM-DAGs, estimate $Z_i = h_i(X_i \mid \mathrm{pa}(X_i))$ of each variable i
- Sample from fitted model to make causal queries

TRAM-DAGs: Experiment 1 (Simulation)

Model fitting: 20,000 training samples, 400 epochs



Sampling from the Fitted TRAM-DAG (L1)

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

- Sample latent value:

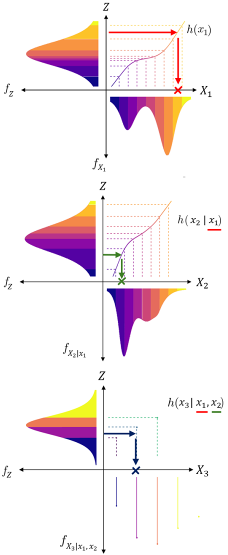
$$z_i \sim F_{Z_i} \quad \text{(e.g., } \texttt{rlogis()} \text{ in R)}$$

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

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

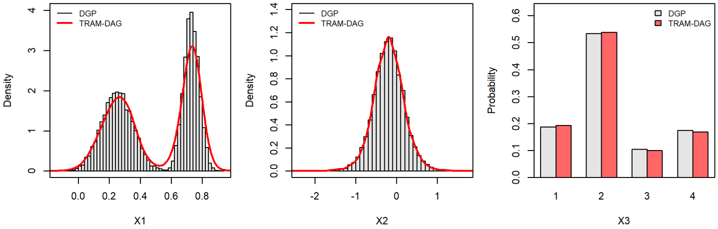
  - **If X_i is ordinal:** find the smallest category x_i such that

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

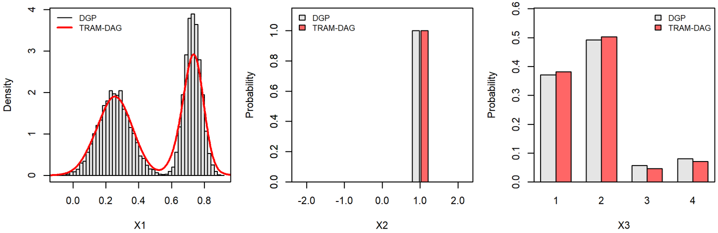


TRAM-DAGs: Experiment 1 (Simulation)

Sampled **Observational** distribution:



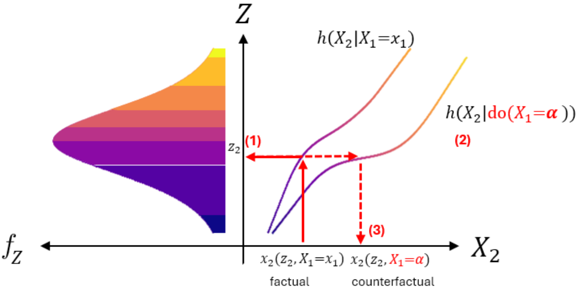
Sampled **Interventional** distribution; do($X_2 = 1$):



Experiment 1: TRAM-DAGs (Simulation)

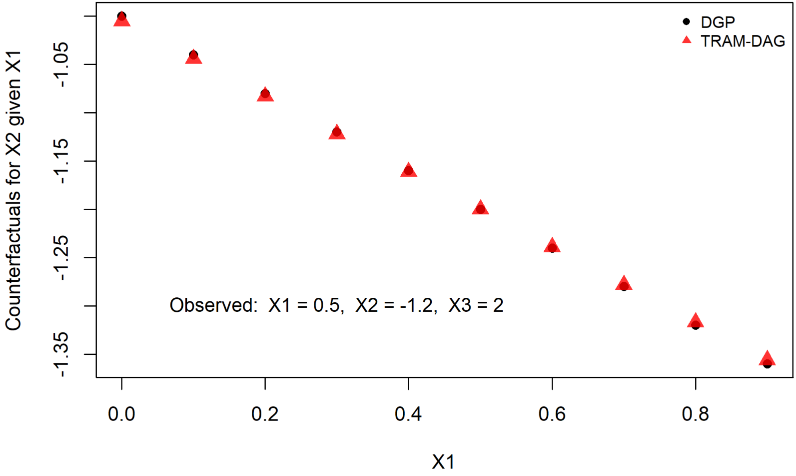
How to determine a counterfactual value for X_2 using Pearl's 3-step procedure (Pearl, 2009):

1. **Abduction**: Infer Z from observed data
2. **Action**: Modify SCM (e.g., do($X = \alpha$))
3. **Prediction**: Infer counterfactual outcome



Experiment 1: TRAM-DAGs (simulation)

Counterfactuals: Counterfactual value of X_2 under varying X_1



Experiment 1: TRAM-DAGs (simulation)

Discussion: With TRAM-DAGs we can

- estimate the functional form of the edges in the DAG
- customize flexibility and interpretability (SI/CI, LS, CS)
- sample from the fitted model (observational/interventional)
- estimate counterfactuals

Individualized Treatment Effects (ITEs)

Individualized Treatment Effect (ITE): Motivation

Why ITE?

- RCTs estimate the Average Treatment Effect (ATE)
- Individuals may respond differently based on covariates

Definition: *Individual treatment effect* (Rubin, 2005)

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

where $Y_i(1)$: outcome if treated, $Y_i(0)$: if not treated

Fundamental problem: We never observe both $Y_i(1)$ and $Y_i(0)$ for the same individual (Holland, 1986).

From Unobservable to Estimable ITE

Goal: Define the *individualized treatment effect (ITE/CATE)* estimand, which we aim to estimate from observed data (Hoogland et al., 2021).

$$
\begin{aligned}
\mathrm{ITE}(\mathbf{x}_i) &= \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X} = \mathbf{x}_i] \\
&= \mathbb{E}[Y_i(1) \mid T = 1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y_i(0) \mid T = 0, \mathbf{X} = \mathbf{x}_i]
\end{aligned}
$$

(by ignorability/exchangeability: no unmeasured confounding)

$$
= \mathbb{E}[Y_i \mid T = 1, \mathbf{X} = \mathbf{x}_i] - \mathbb{E}[Y_i \mid T = 0, \mathbf{X} = \mathbf{x}_i]
$$

(by consistency: observed = potential outcome, e.g. correct label)

Further assumptions:

- **Positivity:** every individual could receive either treatment (e.g. no deterministic assignment)
- **No interference:** one person's treatment does not affect another's outcome

Individualized Treatment Effect (ITE): Models

How did we estimate the potential outcomes $\mathbb{E}[Y_i \mid T = t, \mathbf{X} = \mathbf{x}_i]$?

- **T-learner:**
  1. Fit two separate models on treated and control groups
  2. Predict $\mathbb{E}[Y_i \mid \mathbf{X} = \mathbf{x}_i]$ from each model

  - Logistic regression / Random forest (with hyperparameter tuning)
- **S-learner:**
  1. Fit one model on all data with treatment as a feature
  2. Predict $\mathbb{E}[Y_i \mid \mathrm{do}(T = t), \mathbf{X} = \mathbf{x}_i]$ by setting $T = 0$ and $T = 1$

  - TRAM-DAGs (flexible, interactions, interventions/counterfactuals)

Experiment 2: ITE on International Stroke Trial (IST)

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

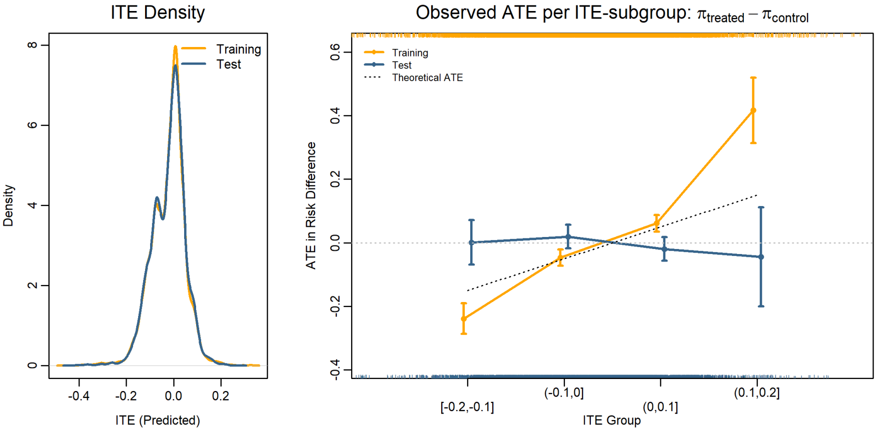
International Stroke Trial (IST):

- Large RCT on stroke patients (19,435 patients, 21 baseline covariates)
- Evaluated the effects of aspirin on death or dependence at 6 months
- Binary treatment and outcome

Research question: Do we reach similar conclusion as Chen et al. (2025) when estimating ITEs with T-learners (logistic regression, tuned random forest) and an S-learner (TRAM-DAGs) on the IST dataset.

Experiment 2: ITE on International Stroke Trial (IST)

Results: with T-learner **tuned random forest** using the `comets` package (Kook, 2024):



Experiment 2: ITE on International Stroke Trial (IST)

Discussion:

- We obtained similar results as Chen et al. (2025)
- Some models suggest moderate treatment effect heterogeneity, but the ITEs do not generalize to the test set (no effect)
- Ground truth is unknown – difficult to determine if no true heterogeneity present or models fail to capture it

Experiment 3: ITE Model Robustness in RCTs (Simulation)

Motivation: ITE estimation did not generalize to the test data on the real-world RCT of the International Stroke Trial (IST). We want to know why!

Research question: What factors contribute to the failure of ITE estimation in causal models?

Setup:

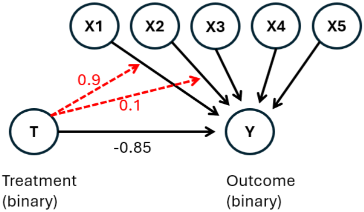
— Simulate different RCT scenarios to understand when ITE estimation fails

— Apply simple model (logistic regression; matching DGP) and non-parametric model (tuned random forest)

Simulation Case 1: Fully Observed

Setup:

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

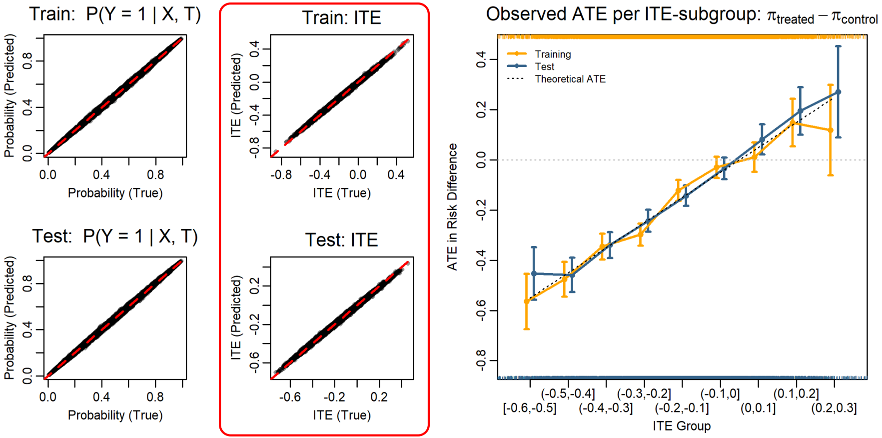


Outcome model:

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

Simulation Case 1: Fully Observed

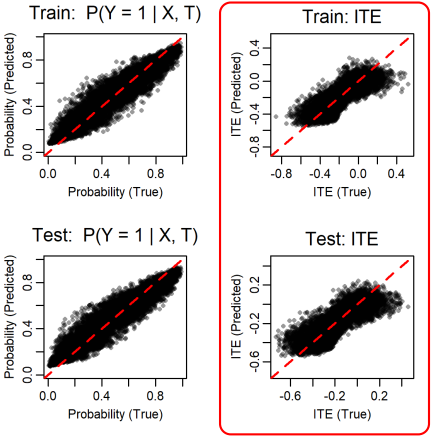
Results with T-learner **logistic regression** (glm):

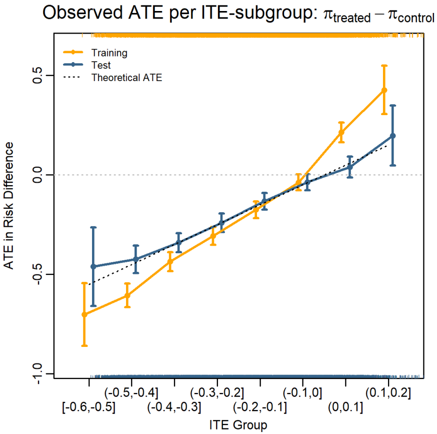


Interpretation: Accurate ITE estimation!

Simulation Case 1: Fully Observed

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

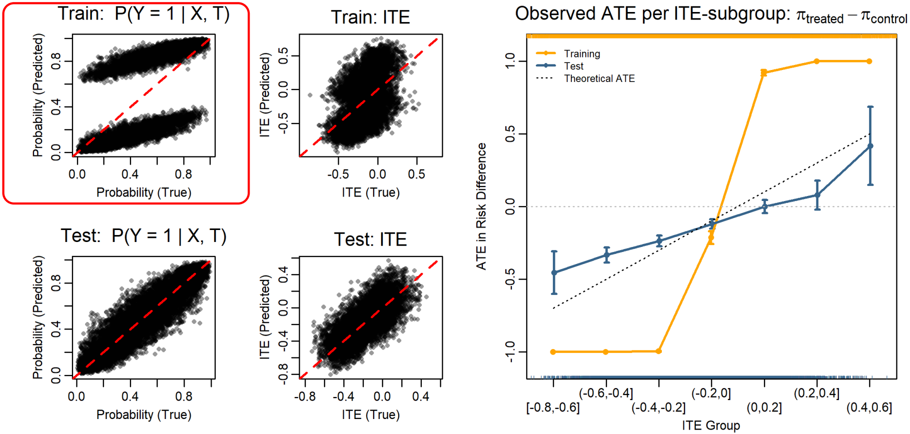




Interpretation: Unbiased ITE estimation!

Simulation Case 1: Fully Observed

Results with **(untuned) T-learner random forest** using the `randomForest` package (Breiman, 2001):

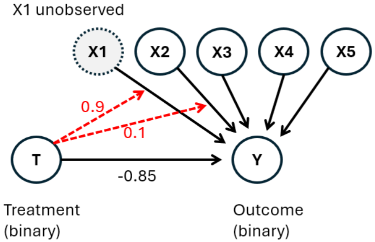


Interpretation: Overfitted, poorly calibrated, leads to worse ITE estimates

Simulation Case 2: Unobserved Interaction

Setup:

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



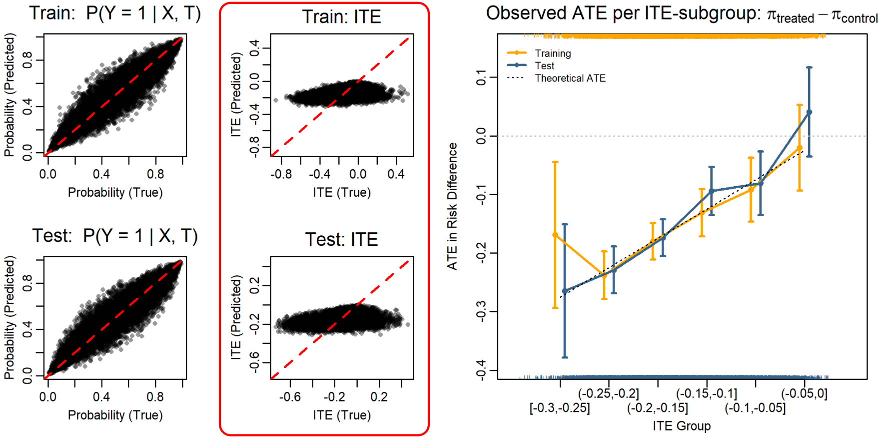
Outcome model:

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

Note: Same DGP, but X_1 is not observed!

Simulation Case 2: Unobserved Interaction

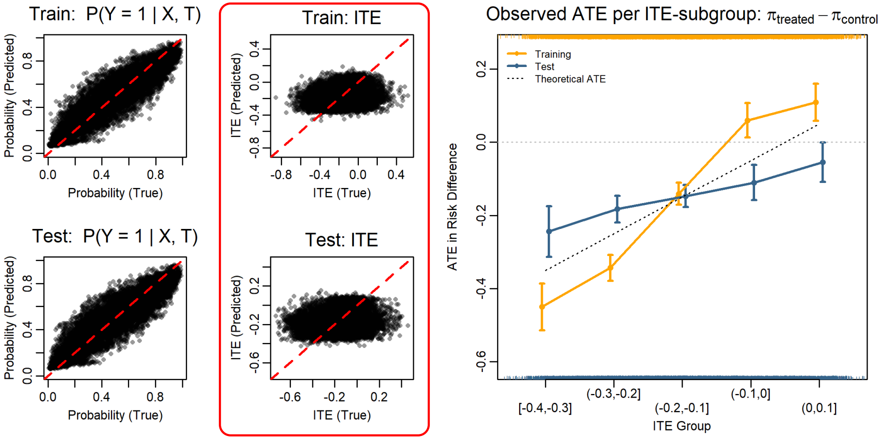
Results with T-learner **logistic regression** (glm):



Interpretation: 1) Model misses positive ITEs, 2) ITE-ATE plot misleading – suggests good calibration, but doesn't detect patients that benefit!

Simulation Case 2: Unobserved Interaction

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

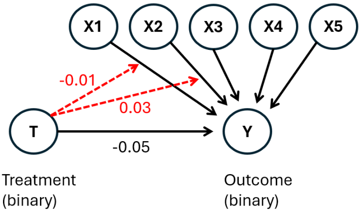


Interpretation: Similar problem as with logistic model!

Simulation Case 3: Fully Observed, Small Effects

Setup:

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



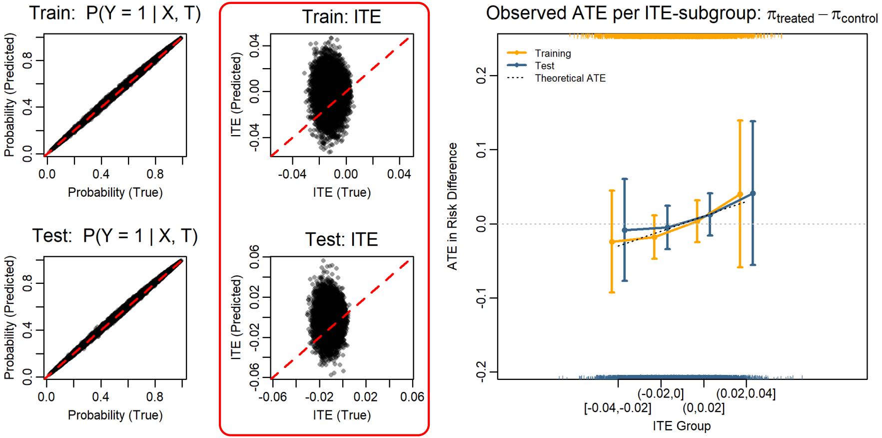
Outcome model:

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

Note: Same DGP, but weak treatment effects!

Simulation Case 3: Fully Observed, Small Effects

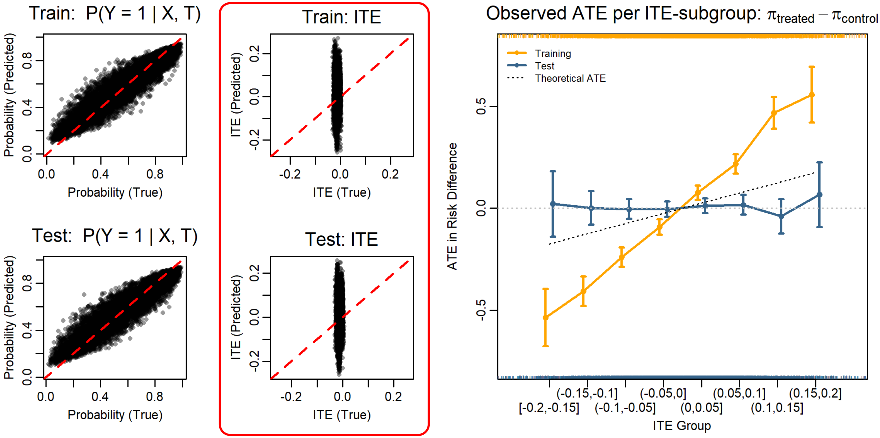
Results with T-learner **logistic regression** (glm):



Interpretation: 1) Predicts too large heterogeneity (model noise?), 2) ITE-ATE plot correctly suggests no significant heterogeneity!

Simulation Case 3: Fully Observed, Small Effects

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



Interpretation: 1) Predicts too large heterogeneity (model noise?), 2) ITE-ATE plot correctly suggests no significant heterogeneity!

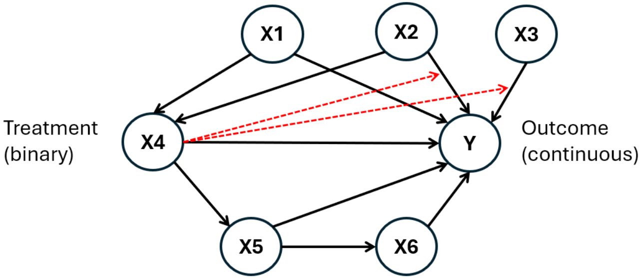
Experiment 3: ITE Model Robustness in RCTs (Simulation)

Key Insights:

- **Calibration** and tuning of models are crucial for reliable ITE estimation
- Ignorability (unconfoundedness) assumption alone may not guarantee unbiased ITEs if important **effect modifiers are unobserved**
- In practice, only the **ITE-ATE plot** is available – it checks ITE calibration in predicted subgroups, but can miss true effect heterogeneity
- **Low true heterogeneity** may be mistaken for model failure

These factors may explain the limited ITE performance in the IST dataset.

Experiment 4: ITE Estimation with TRAM-DAGs



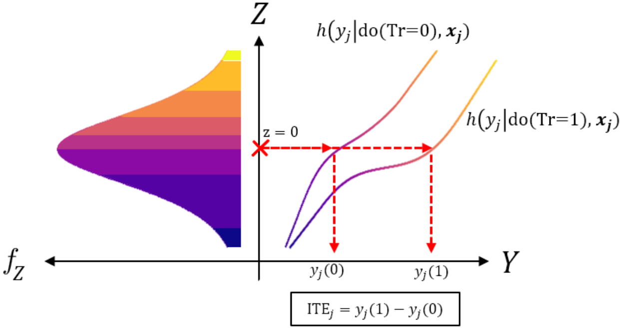
DGP:

- $X_1, X_2, X_3 \sim \mathcal{N}(\mathbf{0}, \Sigma)$
- X_4 (treatment) depends probabilistically on X_1 and X_2 via a logistic model
- $X_5 = h_5^{-1}(Z_5 - 0.8\, X_4) \quad \rightarrow$ (depends on treatment)
- $X_6 = h_6^{-1}(Z_6 + 0.5\, X_5) \quad \rightarrow$ (depends on treatment through X_5)
- $Y = h_7^{-1}(Z_7 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \beta_4 X_4 - \beta_5 X_5 - \beta_6 X_6 - X_4 \cdot (\beta_{2,Tr} X_2 + \beta_{3,Tr} X_3))$

Experiment 4: ITE Estimation with TRAM-DAGs

We define the ITE as the difference in medians of potential outcomes:

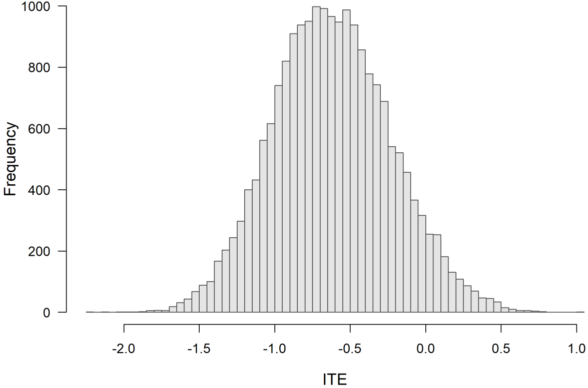
$$\text{ITE} = \text{median}(Y \mid \text{do}(T = 1), \mathbf{X}) - \text{median}(Y \mid \text{do}(T = 0), \mathbf{X})$$



Experiment 4: ITE Estimation with TRAM-DAGs

Resulting ITEs from the DGP in terms of difference in medians of potential outcomes:

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

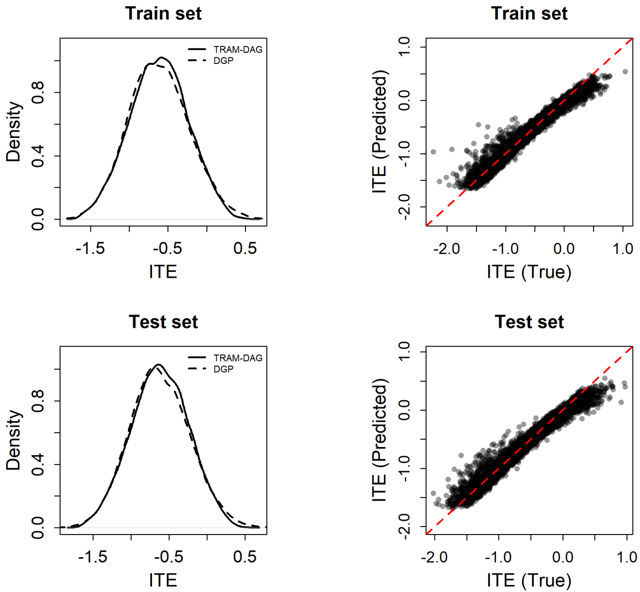


Experiment 4: ITE Estimation with TRAM-DAGs

Estimate ITEs with TRAM-DAGs (S-learner approach) from observed data:

1. Fit the TRAM-DAG on the training set (fully flexible – CI – to allow for interactions)
2. Compute potential outcomes as $\text{median}(Y \mid \text{do}(T = t), \mathbf{X}_t)$ for $t \in \{0, 1\}$
3. $\text{ITE} = \text{median}(Y \mid \text{do}(T = 1), \mathbf{X}_1) - \text{median}(Y \mid \text{do}(T = 0), \mathbf{X}_0)$

ITE Estimation with TRAM-DAGs (Results)



Key Findings

Findings: TRAM-DAGs

- Customizable; accurately recovers causal relationships in known DAG; allows sampling of L1-L3
- Can model interactions between variables

Findings: Individualized treatment effects (ITE)

- Calibration is important for ITE prediction
- Missing effect modifiers (or weak heterogeneity) are problematic
- TRAM-DAGs yield unbiased ITEs when DAG is correct and heterogeneity exists

Outlook

Limitations

- Simulations may not reflect real-world complexity
- TRAM-DAGs are computationally expensive (long training time)
- TRAM-DAGs require correct model specification for interpretability
- ITE estimation for continuous outcomes used medians of potential outcomes instead of expected values

Recommendations

- Apply TRAM-DAGs to real-world datasets, including semi-structured data
- Investigate ITE estimation under unobserved effect modifiers

References I

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

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. arXiv preprint 2501.04061.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., E. Harrell Jr, F., Moons, K. G. M., Debray, T. P. A., and Reitsma, J. B. (2021). A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*, 40(26):5961–5981.

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.

References II

Kook, L. (2024). *comets: Covariance Measure Tests for Conditional Independence*. R package version 0.1-1.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition.

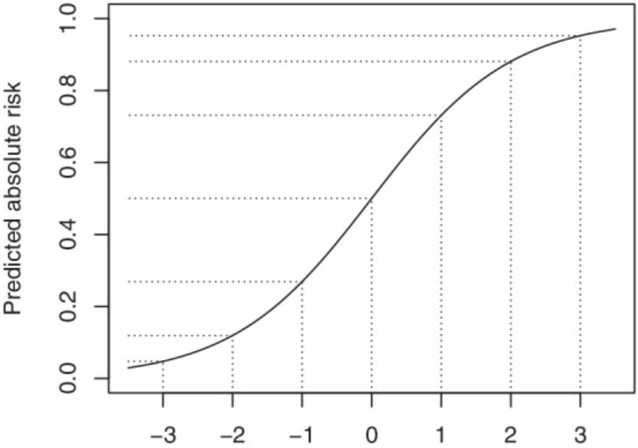
Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Sick, B. and Dürr, O. (2025). Interpretable neural causal models with TRAM-DAGs. arXiv preprint 2503.16206, accepted at the CLeaR 2025 Conference.

Sick, B., Hothorn, 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.

Heterogeneity

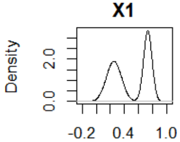
Heterogeneity despite no interaction effects in logistic model (Hoogland et al., 2021).



TRAM-DAGs: Experiment 1 (simulation)

Data-generating process (DGP):

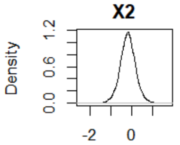
X_1**:** Continuous, bimodal. *Source node* (independent).



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

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

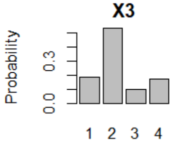
$$\boxed{h(X_2 \mid X_1) = h_I(X_2) + \beta_{12} X_1}$$



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

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

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



TRAM-DAGs: Experiment 1 (simulation)

Construct Model: Modular Neural Network

Inputs: Observations + assumed structure

Outputs:

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

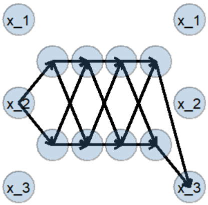
Assemble transformation functions:

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

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

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

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



CS_{X_2} on X_3