Interpretable Causal Inference for High-Stakes Decisions

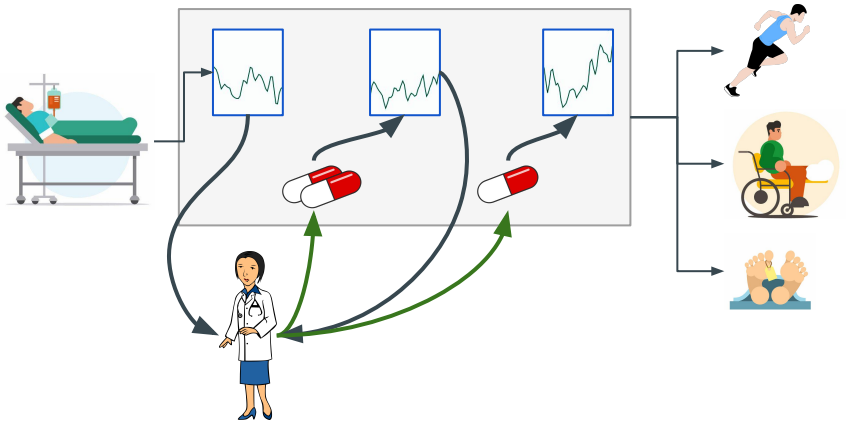
Harsh Parikh
Dept. of Computer Science
Duke University

Duke | ALMOST MATCHING EXACTLY LAB

Bills' Damar Hamlin in critical condition after collapse on field

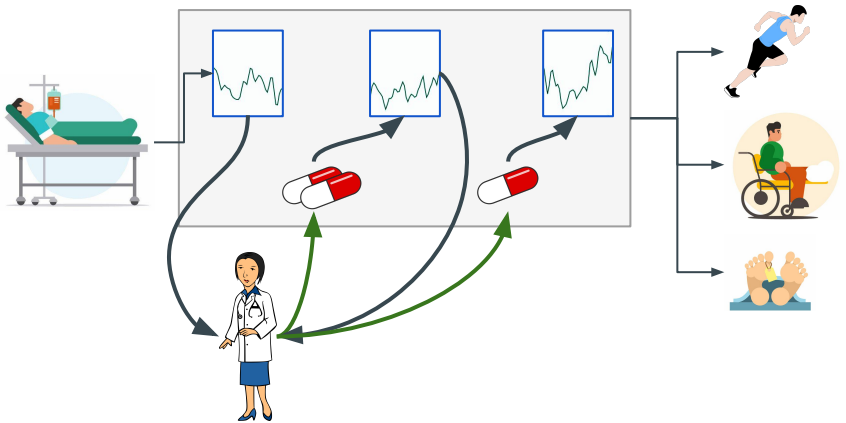
Nation Jan 3, 2023 9:18 AM EST

Treating Patients in ICU



- **Which** drugs to use?
- **When** to administer these drugs?
- **What** should be the drug **dose**?

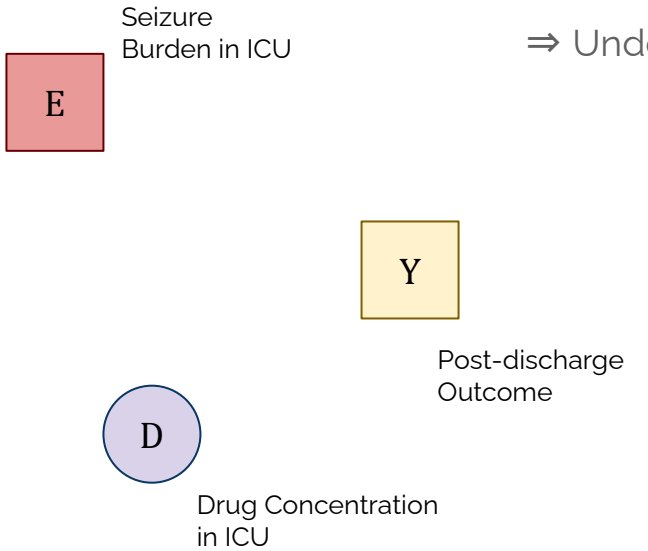
Treating Patients in ICU



- **Which** drugs to use?
- **When** to administer these drugs?
- **What** should be the drug **dose**?

Doctors need to understand the **risk-reward trade-offs**!

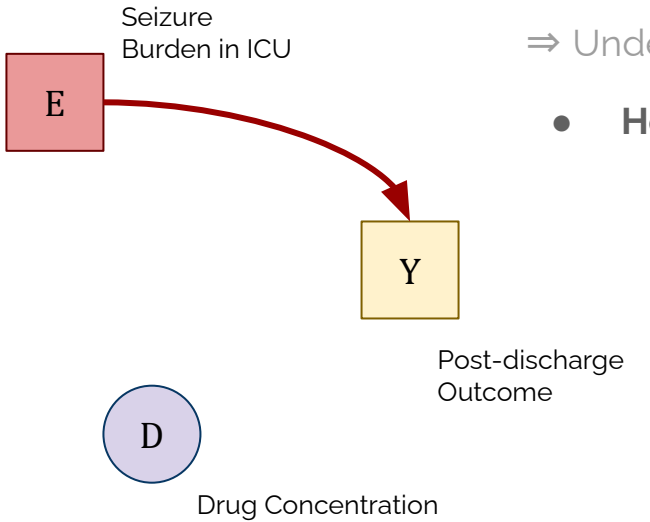
How to Treat Seizures in Critically-ill Patients?



Understanding **risk-reward trade-offs**

⇒ Understanding **causal dependencies**

How to Treat Seizures in Critically-ill Patients?

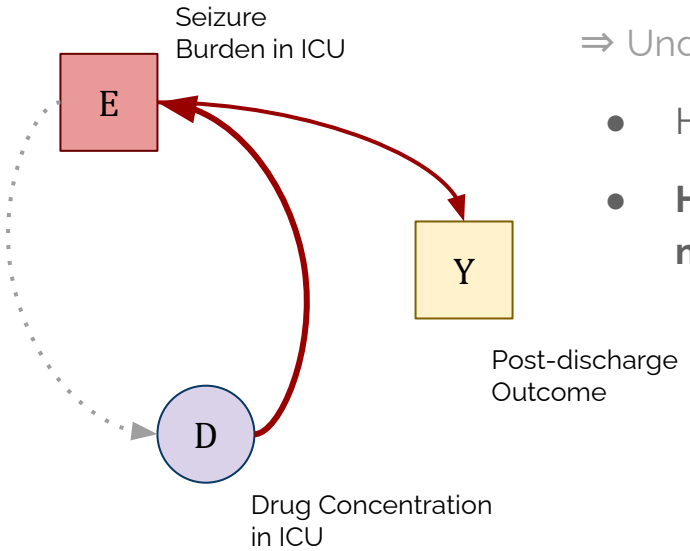


Understanding **risk-reward trade-offs**

⇒ Understanding **causal dependencies**

- **How bad are seizures for brain?**

How to Treat Seizures in Critically-ill Patients?

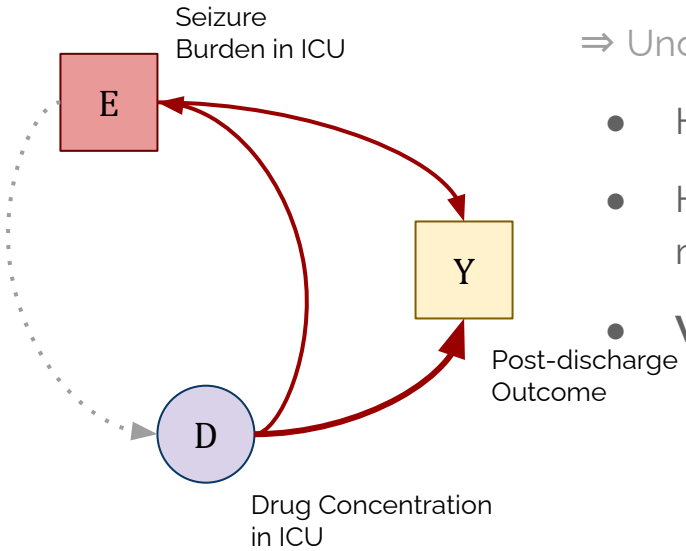


Understanding **risk-reward trade-offs**

⇒ Understanding **causal dependencies**

- How bad are seizures for brain?
- **How effective are anti-seizure medications (ASMs)?**

How to Treat Seizures in Critically-ill Patients?

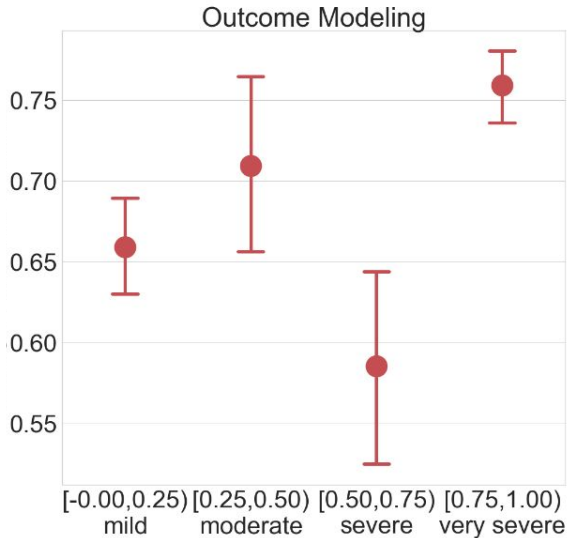


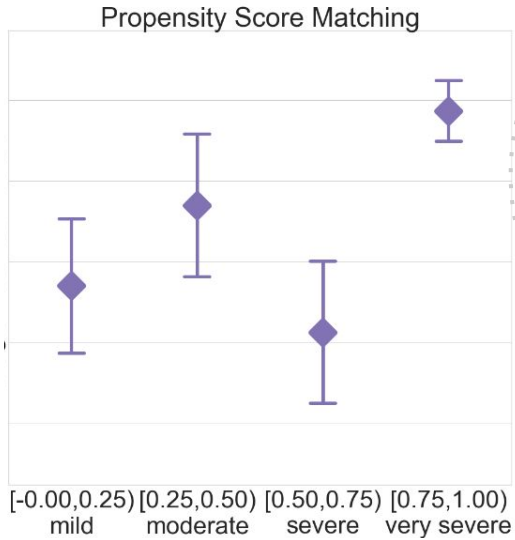
Understanding **risk-reward trade-offs**

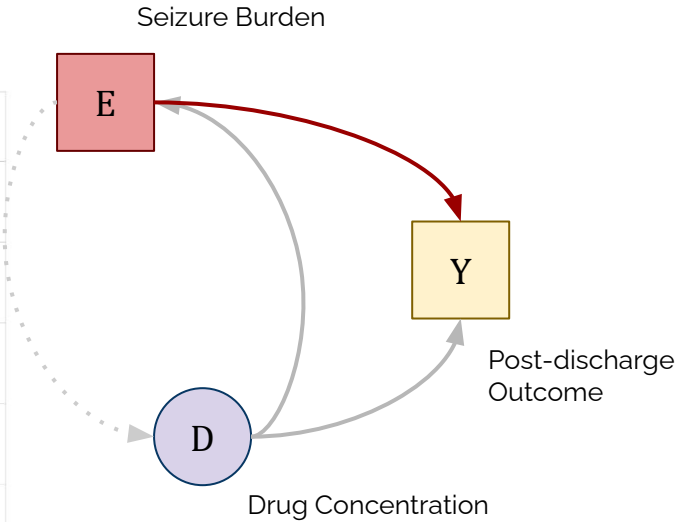
⇒ Understanding **causal dependencies**

- How bad are seizures for brain?
- How effective are anti-seizure medications (ASMs)?
- **What are long-term effects of ASMs?**

How Bad are Seizures for Brain? (Results from the Literature)

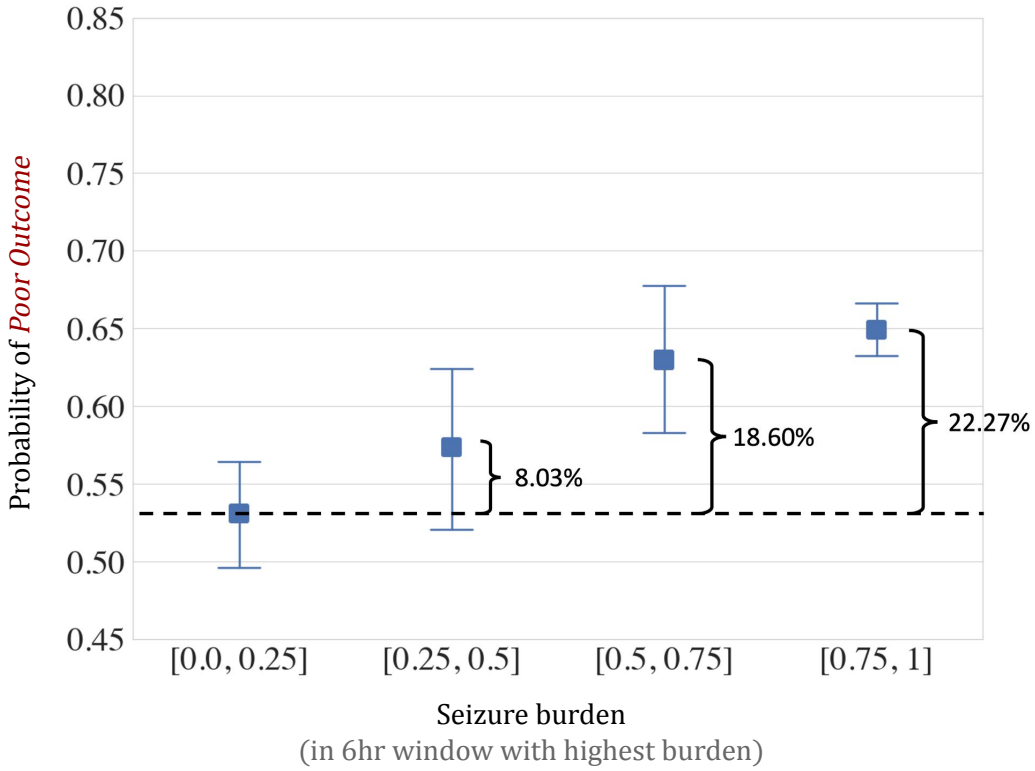


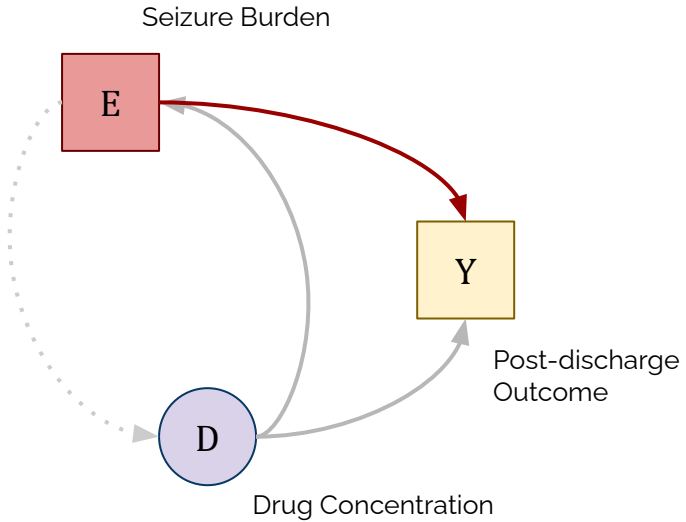




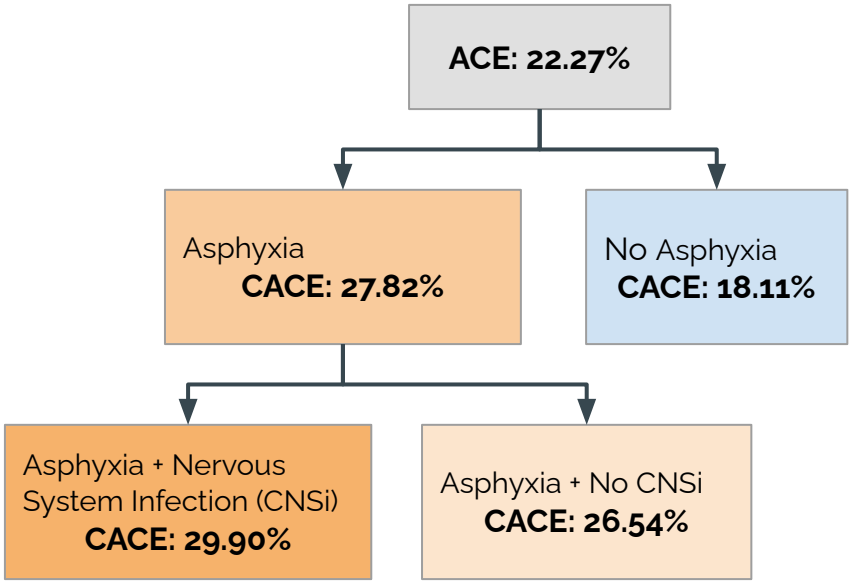
1. Payne *et al*, 2014
2. De Marchis *et al*, 2016
3. Zafar *et al*, 2018

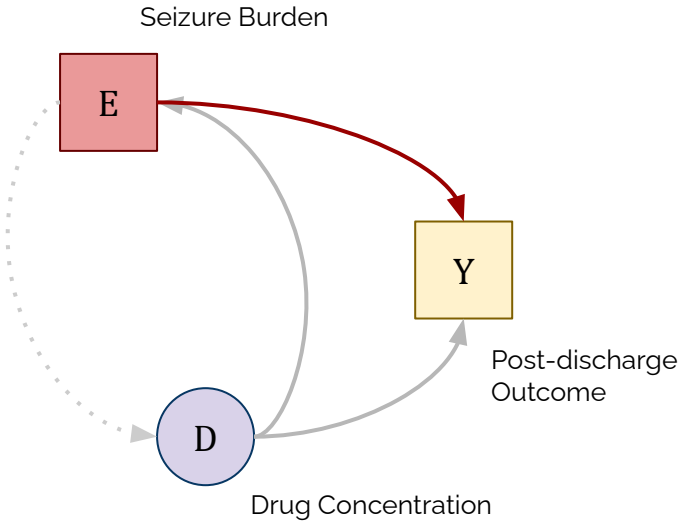
How Bad are Seizures for Brain? (Our Analysis)





How Bad are Seizures for Brain? (Heterogeneity)

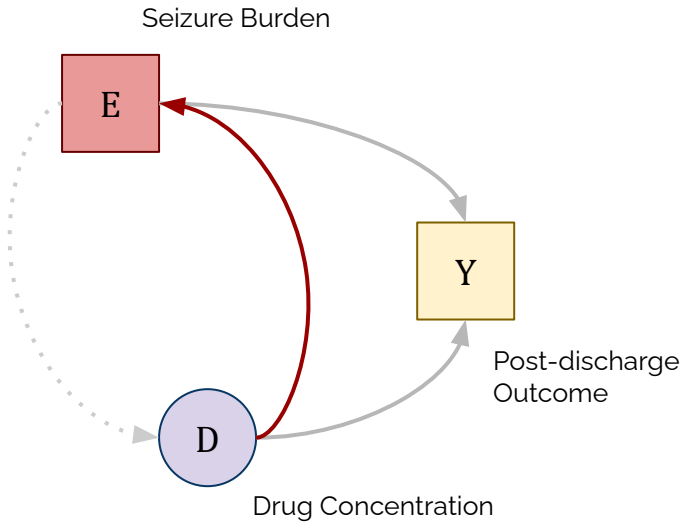




How Effective are ASMs in Reducing Seizure Burden?

Parameters of Interest

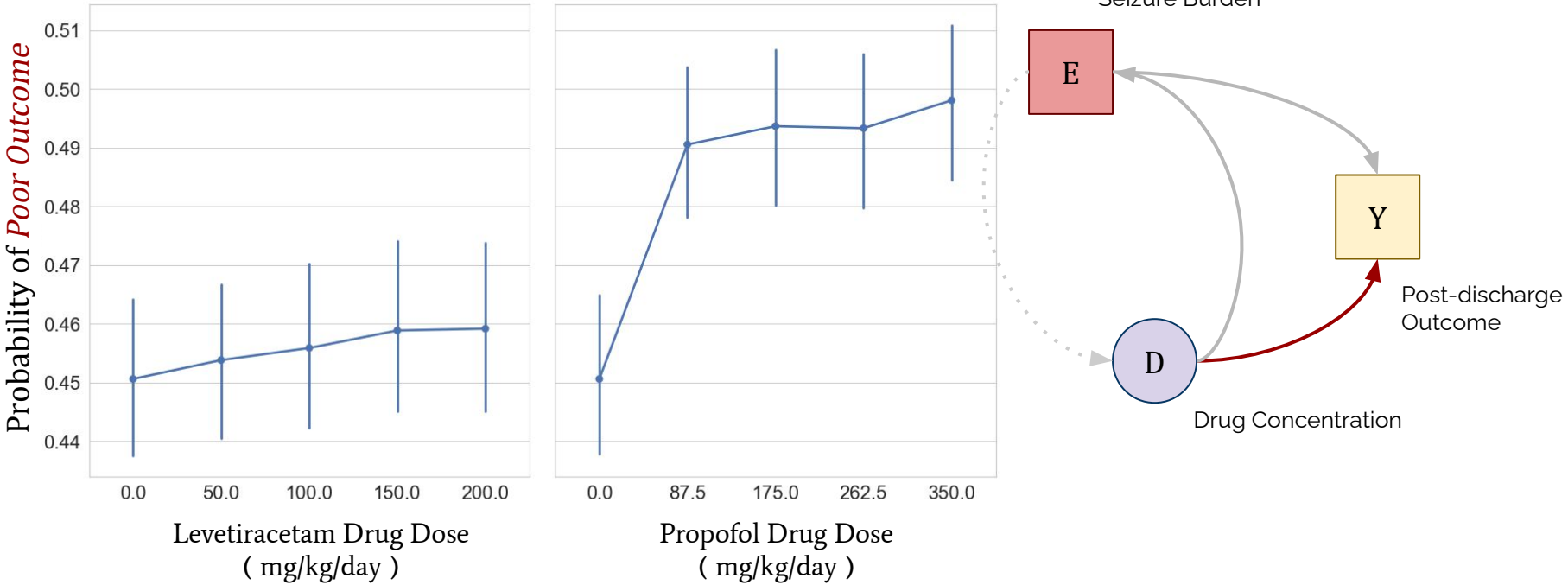
- **ED50:** Dose required to reduce Sz burden by 50%
- α: binding between the receptor and the ligand (Hill's coefficient)



Results: How Effective are ASMs in Reducing Seizure Burden?

	Parameter Description	Levetiracetam	Propofol
Type		*Non Sedating*	*Sedating*
Avg. ED50	Dose required to reduce Seizure burden by 50%	5.57 mg/kg	1.88 mg/kg
Avg. α	Binding between the receptor and the ligand (Hill's coefficient)	1.168	2.34

Results: What are Long-term Negative Effects of ASMs?



Results: Implications

- Intense burst of seizures even for a short-period can lead to poor outcome
    - Design anti-seizure medication (ASM) treatment regime to suppress these intense seizure bursts

Results: Implications

- Intense burst of seizures even for a short-period can lead to poor outcome
    - Design anti-seizure medication (ASM) treatment regime to suppress these intense seizure bursts
- Patients with asphyxia and CNS infection are adversely affected by seizures
    - Prioritize aggressive treatment for patients with these diagnosis

Results: Implications

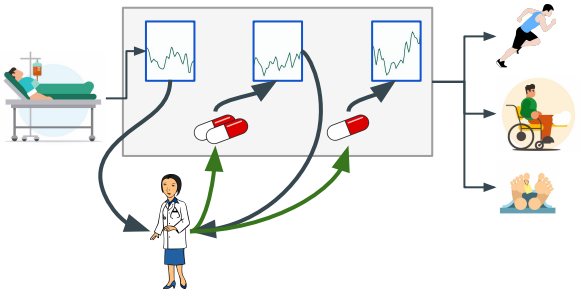
- Intense burst of seizures even for a short-period can lead to poor outcome
  - Design anti-seizure medication (ASM) treatment regime to suppress these intense seizure bursts
- Patients with asphyxia and CNS infection are adversely affected by seizures
  - Prioritize aggressive treatment for patients with these diagnosis
- Primary seizure management using levetiracetam
  - Use of propofol limited only to control seizure bursts

Results: Implications

- Intense burst of seizures even for a short-period can lead to poor outcome
  - Design anti-seizure medication (ASM) treatment regime to suppress these intense seizure bursts
- Patients with asphyxia and CNS infection are adversely affected by seizures
  - Prioritize aggressive treatment for patients with these diagnosis
- Primary seizure management using levetiracetam
  - Use of propofol limited only to control seizure bursts
- Multi-center observational and experimental studies to understand effect heterogeneity across patients and ensure generalizability

Causal Methods for Complex Data

Research Overview



Causal Methods for Complex Data

Auditable Methods

- **P**RV, JMLR
- **P**RV, Obs Study
- L**P**VRP, *Submitted*
- K**P**VRSW, AAAI W3PHIAI

Mechanistic Model Augmented Study Design

- **P***H*S*ZVRW, *Submitted*, Lancet Digital Health
- **P**ZSVRW, *working*

Data Fusion for Generalizability and Efficiency

- M*O***P***VRR, *Working*

Interpretable RL for Optimal Treatment Regimes

- **P***S*VRW, *working*

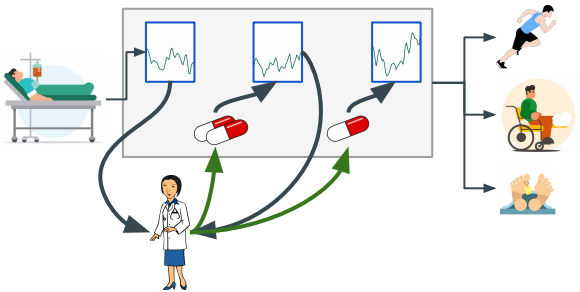
Causal Methods for Relational Data

- S**P**KRGS, SIGMOD
- **P**GSR, *Preprint*

Validating Causal Inference Methods

- **P**VXT, ICML

Research Overview



Causal Methods for Complex Data

Auditable Methods

- **P**RV, JMLR
- **P**RV, Obs Study
- L**P**VRP, *Submitted*
- K**P**VRSW, AAAI W3PHIAI

Mechanistic Model Augmented Study Design

- **P***H*S*ZVRW, *Submitted*, Lancet Digital Health
- **P**ZSVRW, *working*

Validating Causal Inference Methods

- **P**VXT, ICML

Causal Methods for Relational Data

- S**P**KRGS, SIGMOD
- **P**GSR, *Preprint*

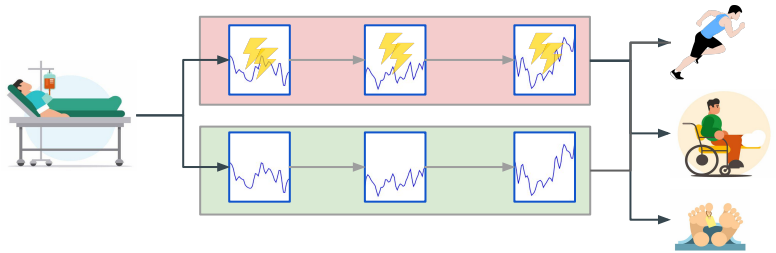
Data Fusion for Generalizability and Efficiency

- M*O***P***VRR, *Working*

Interpretable RL for Optimal Treatment Regimes

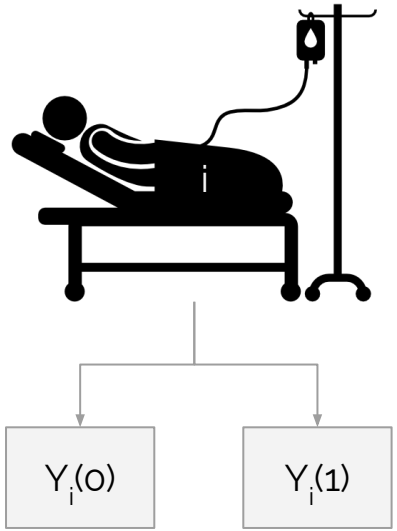
- **P***S*VRW, *working*

Notations: What-ifs, Potential Outcomes and Causal Effects



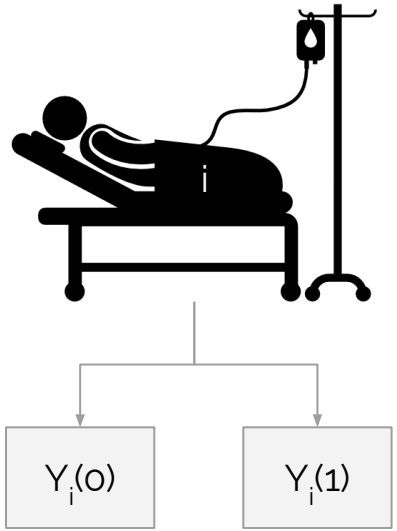
Notations: What-ifs, Potential Outcomes and Causal Effects

	Seizure=0	Seizure=1
Potential Outcomes	$Y_i(0)$	$Y_i(1)$



Notations: What-ifs, Potential Outcomes and Causal Effects

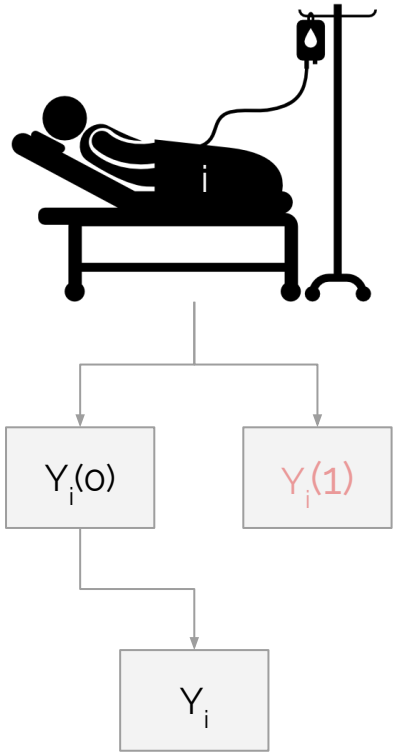
	Seizure=0	Seizure=1
Potential Outcomes	$Y_i(0)$	$Y_i(1)$



Causal Effect of Seizures = $Y_i(1) - Y_i(0)$

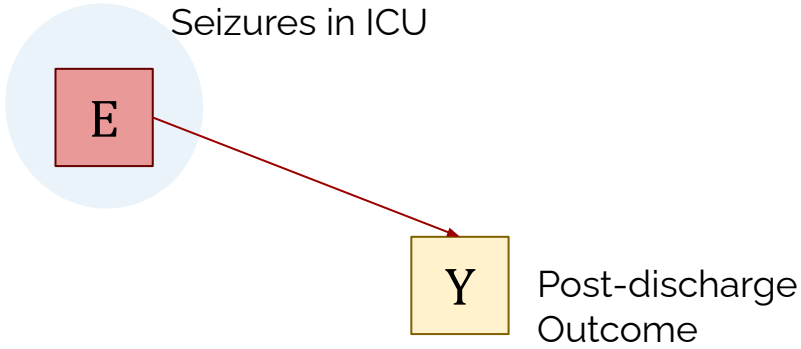
Notations: What-ifs, Potential Outcomes and Causal Effects

	Seizure=0	Seizure=1
Potential Outcomes	$Y_i(0)$	$Y_i(1)$

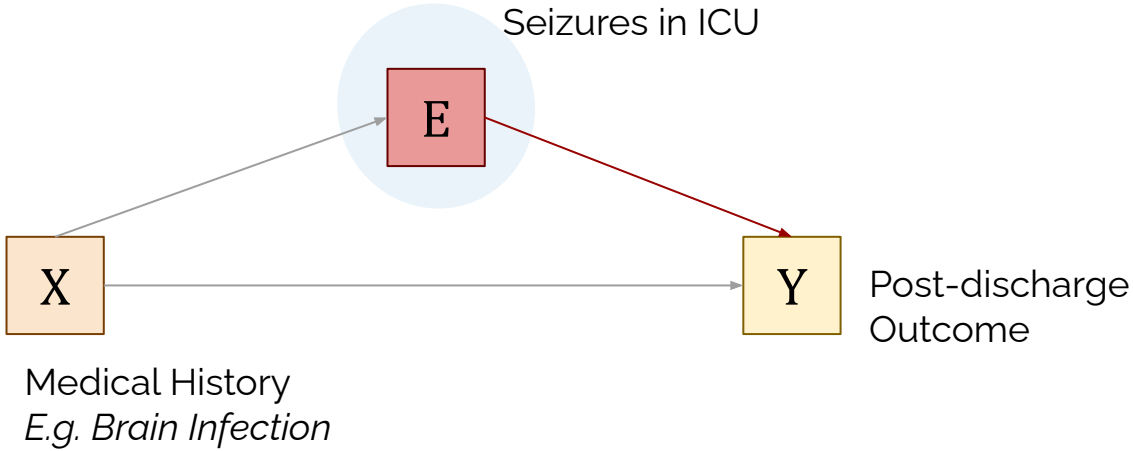


Causal Effect of Seizures = $Y_i(1) - Y_i(0) = ? - Y_i$

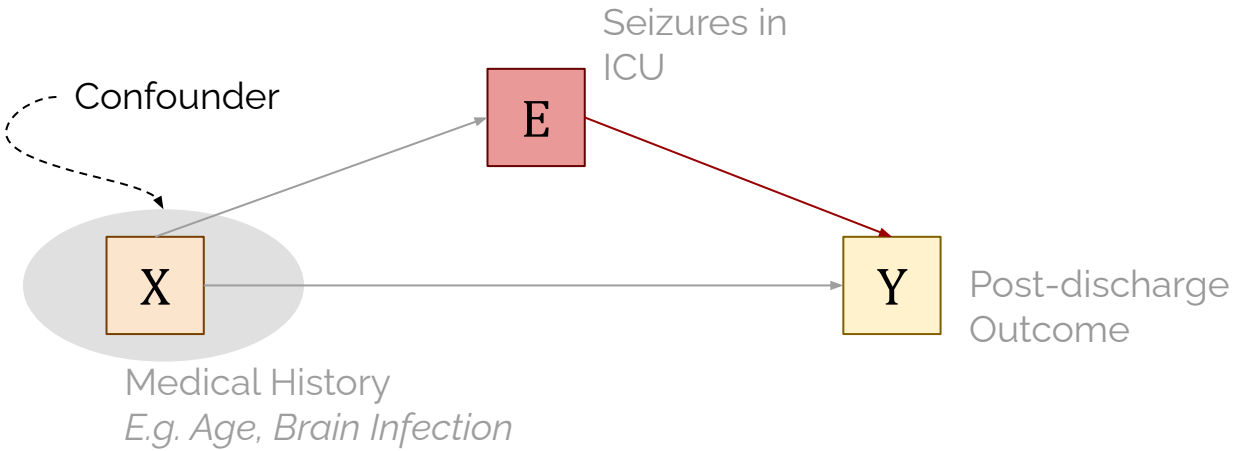
Confounding



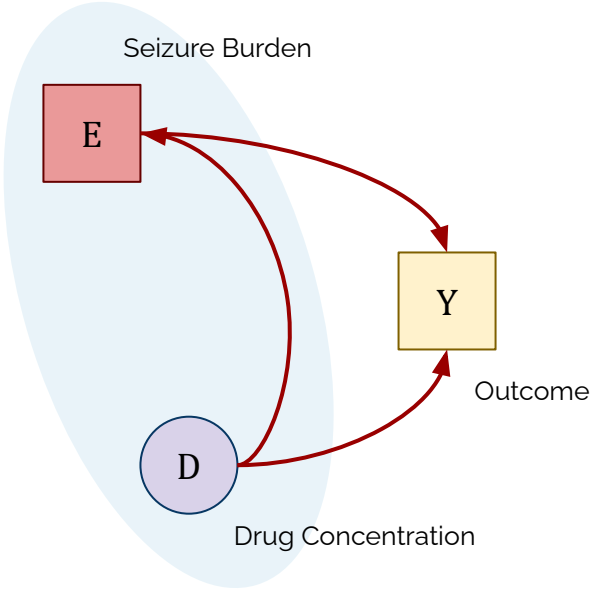
Confounding



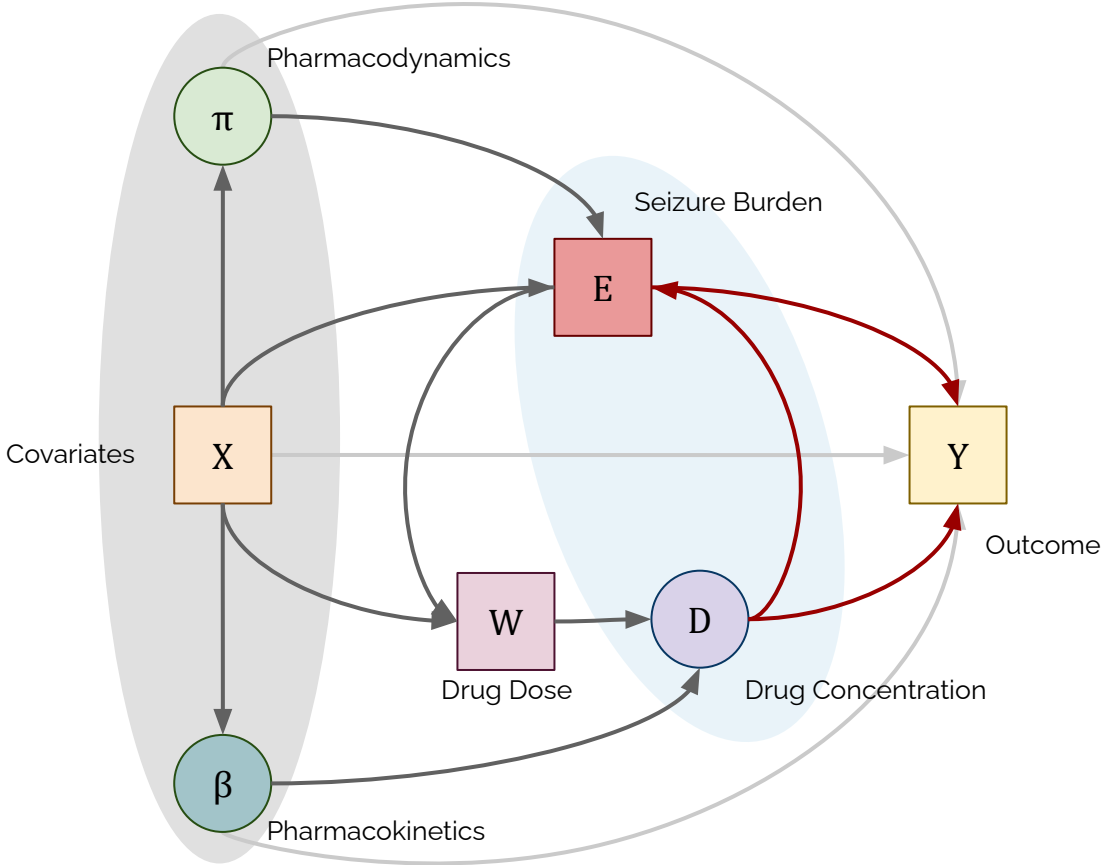
Confounding



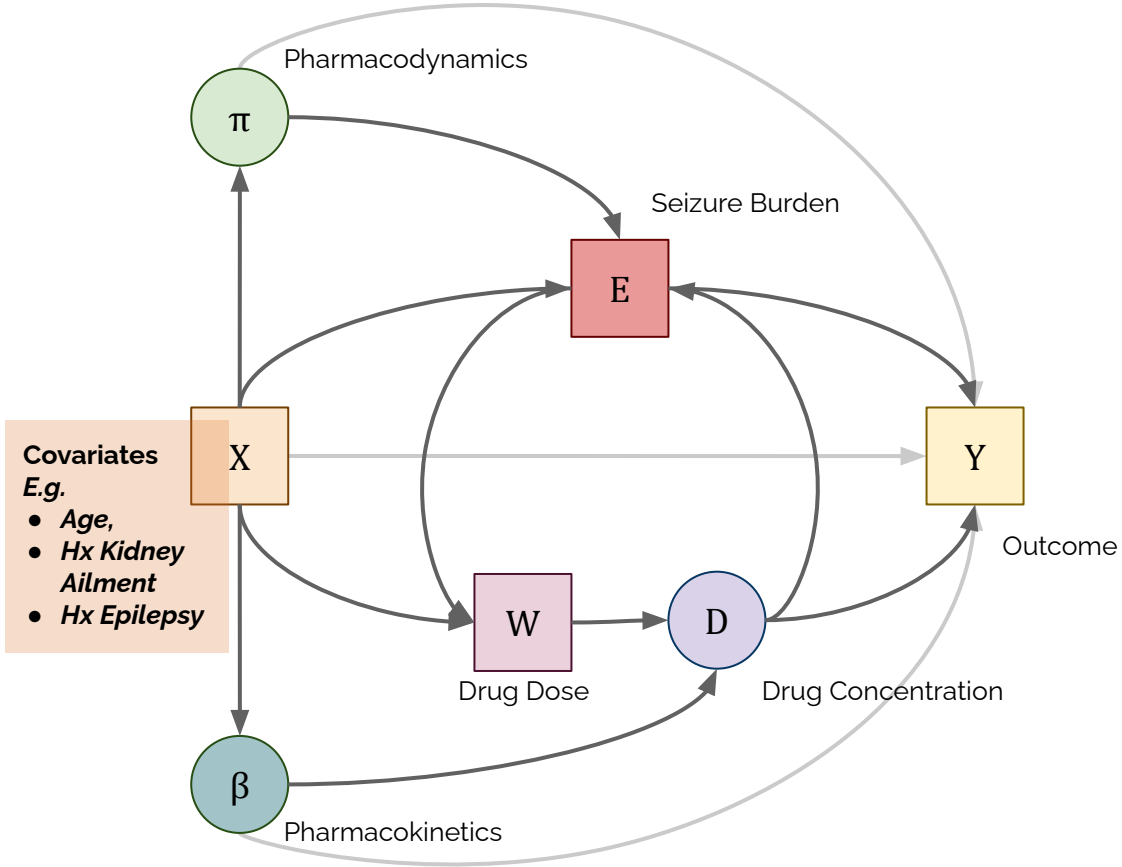
How to Treat Seizures in Critically-ill Patients?



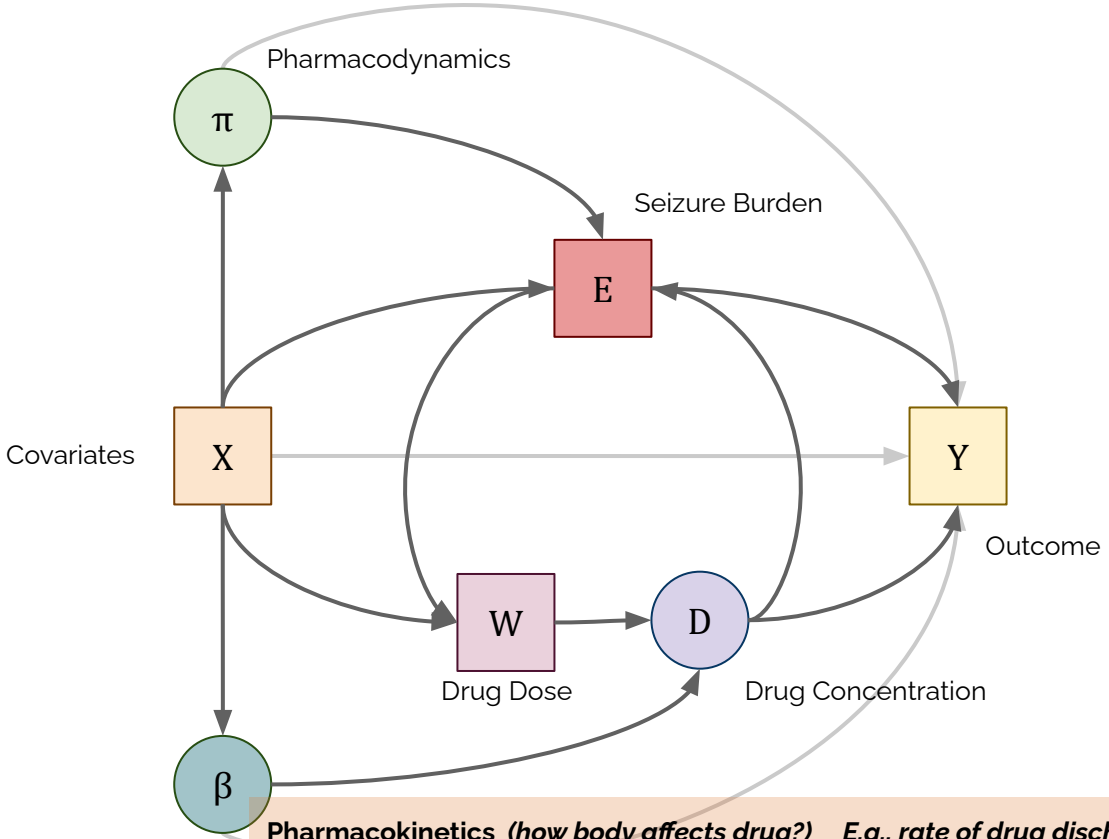
How to Treat Seizures in Critically-ill Patients?



How to Treat Seizures in Critically-ill Patients?

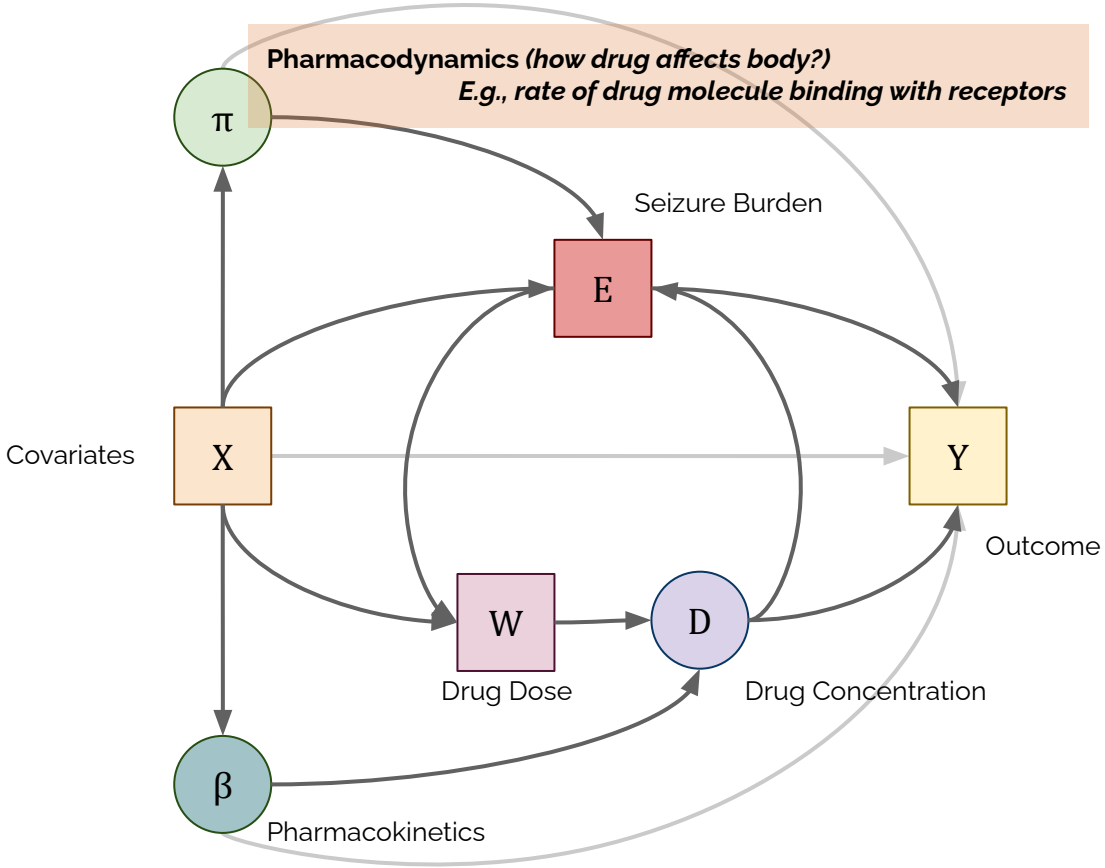


How to Treat Seizures in Critically-ill Patients?



Pharmacokinetics ***(how body affects drug?)*** ***E.g., rate of drug discharge/excretion***

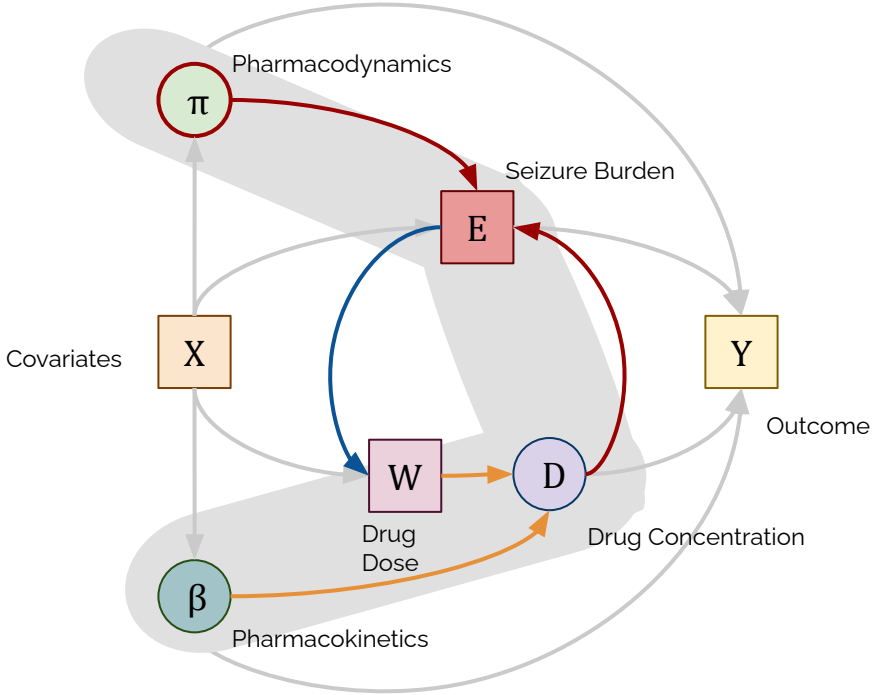
How to Treat Seizures in Critically-ill Patients?



Pharmacology and Entangled Exposures

$$E_t \sim f(E_{0:t-1}, D_t, X, \pi)$$

$$D_t \sim g(W_{0:t}, \beta)$$



Pharmacology and Entangled Exposures

Past Seizure Burden

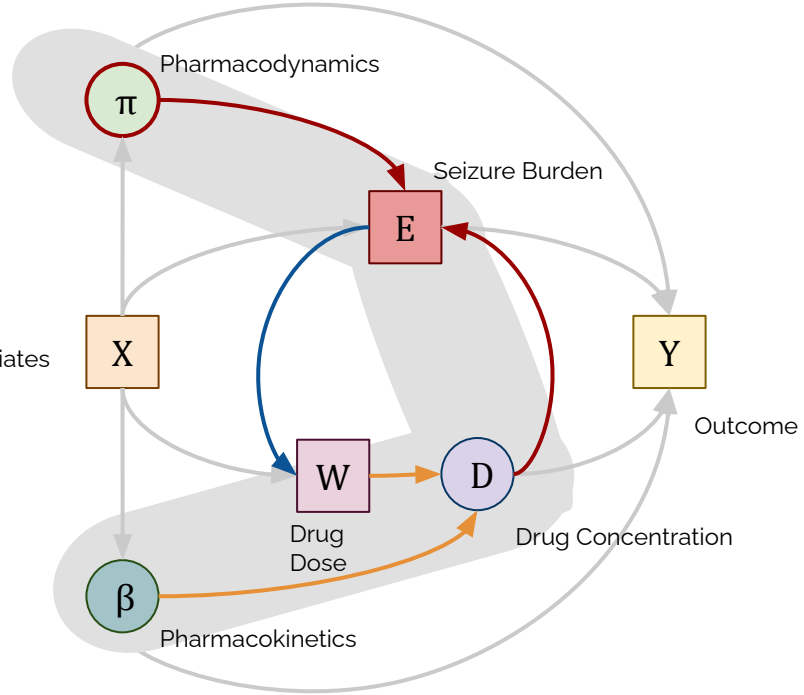
Current Drug Concentration

$$E_t \sim f(E_{0:t-1}, D_t, X, \pi)$$

Covariates

Pharmacodynamic Parameters

$$D_t \sim g(W_{0:t}, \beta)$$



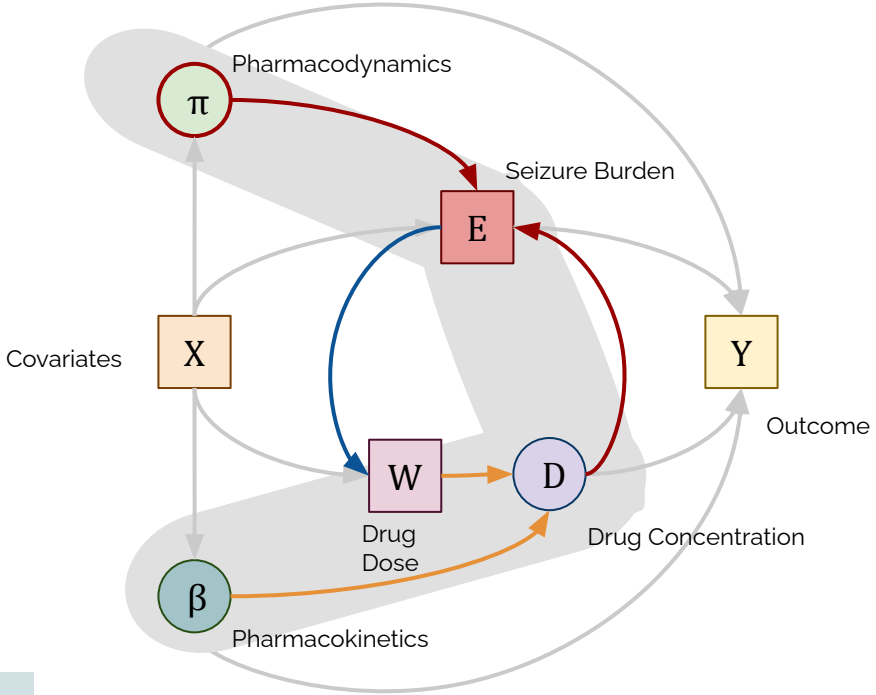
Pharmacology and Entangled Exposures

$$E_t \sim f(E_{0:t-1}, D_t, X, \pi)$$

$$D_t \sim g(W_{0:t}, \beta)$$

Past and Current Drugs Administered

Pharmacokinetic Parameter

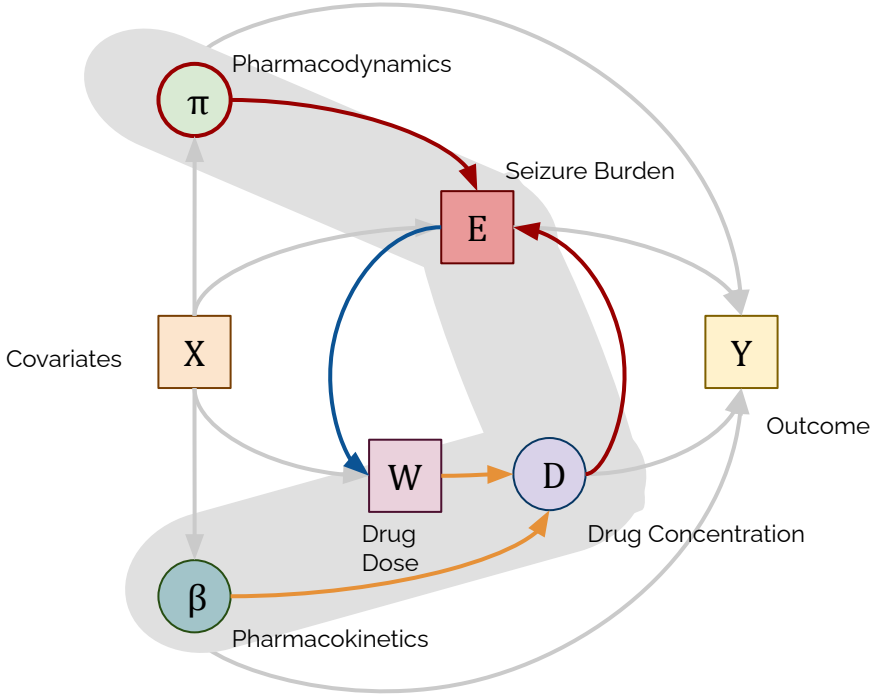


Mechanistic Modeling of Pharmacology

$$E_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{\alpha_{i,j}}}{D_{i,t,j}^{\alpha_{i,j}} + ED_{50,i,j}^{\alpha_{i,j}}}$$

$$\pi_i = \{ED_{50,i,j}; \alpha_{i,j}\}_j$$

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\beta_j} D_{i,t,j} + W_{i,t,j}$$



Mechanistic Modeling of Pharmacology

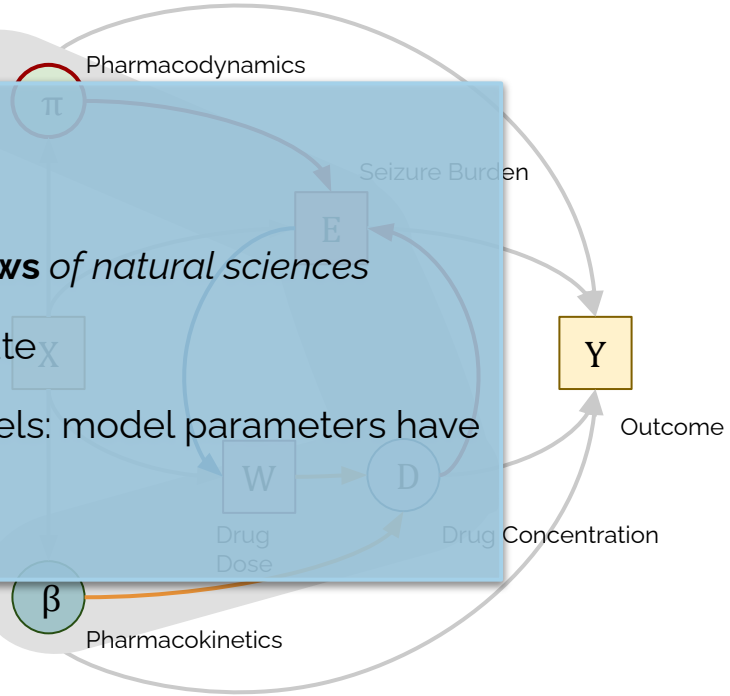
$$E_{i,t} = 1 - \sum_j \frac{D$$

$$\pi_i = \{E$$

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\beta_j} D_{i,t,j} + W_{i,t,j}$$

Why Mechanistic model?

- Based on the **fundamental laws** *of natural sciences*
- **Lesser data** needed to calibrate
- Mechanistic vs statistical models: model parameters have an *actual* **physical meaning**

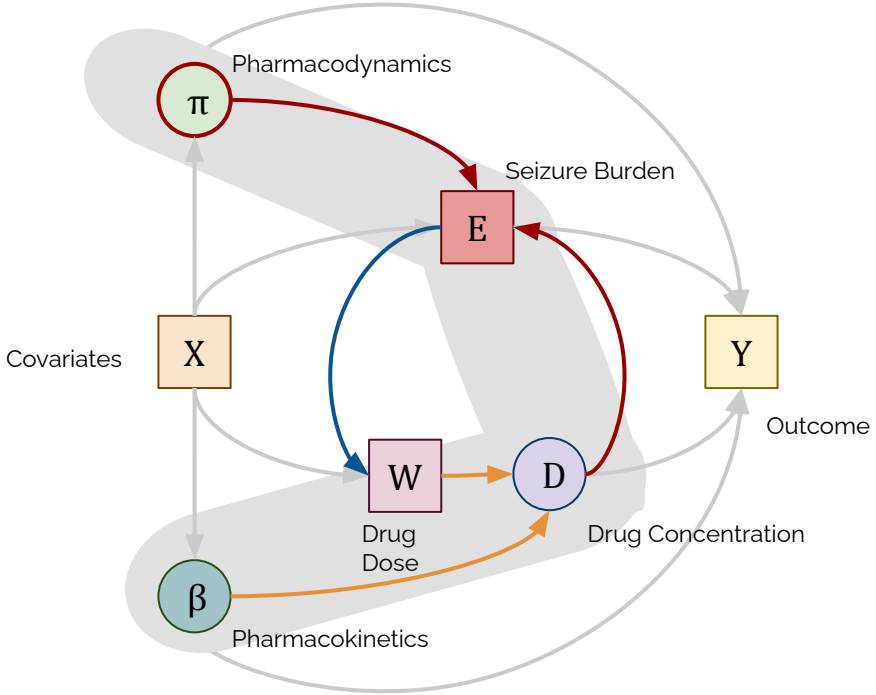


Mechanistic Modeling of Pharmacology

$$E_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{\alpha_{i,j}}}{D_{i,t,j}^{\alpha_{i,j}} + ED_{50,i,j}^{\alpha_{i,j}}}$$

$$\pi_i = \{ED_{50,i,j}; \alpha_{i,j}\}_j$$

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\beta_j} D_{i,t,j} + W_{i,t,j}$$



Mechanistic Modeling of Pharmacology

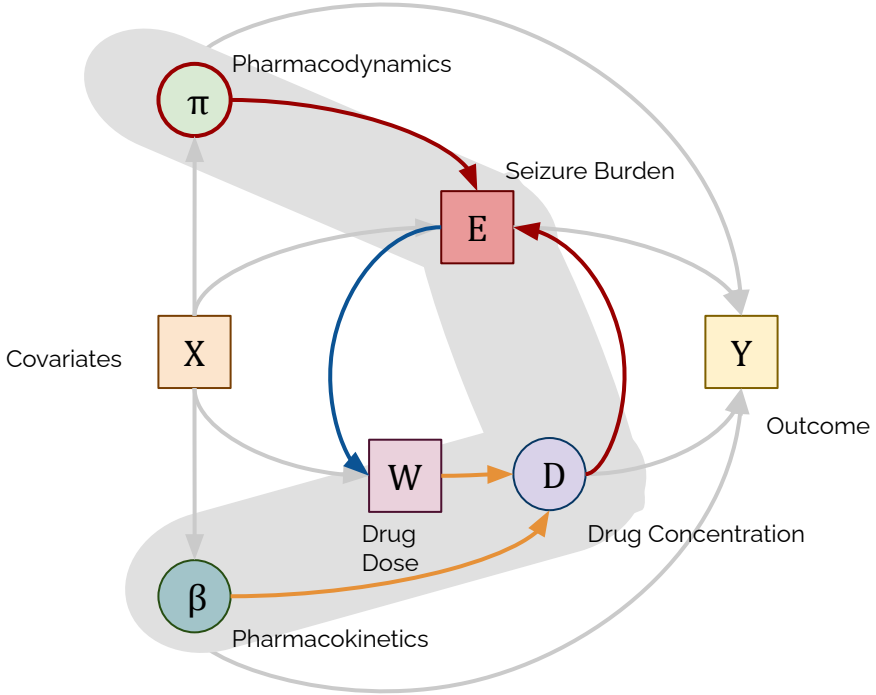
$$E_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{\alpha_{i,j}}}{D_{i,t,j}^{\alpha_{i,j}} + ED_{50,i,j}^{\alpha_{i,j}}}$$

Drug dose for 50% Sz burden

Hill's coefficient / Ligand-receptor affinity

$$\pi_i = \{ED_{50,i,j}; \alpha_{i,j}\}_j$$

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\beta_j} D_{i,t,j} + W_{i,t,j}$$



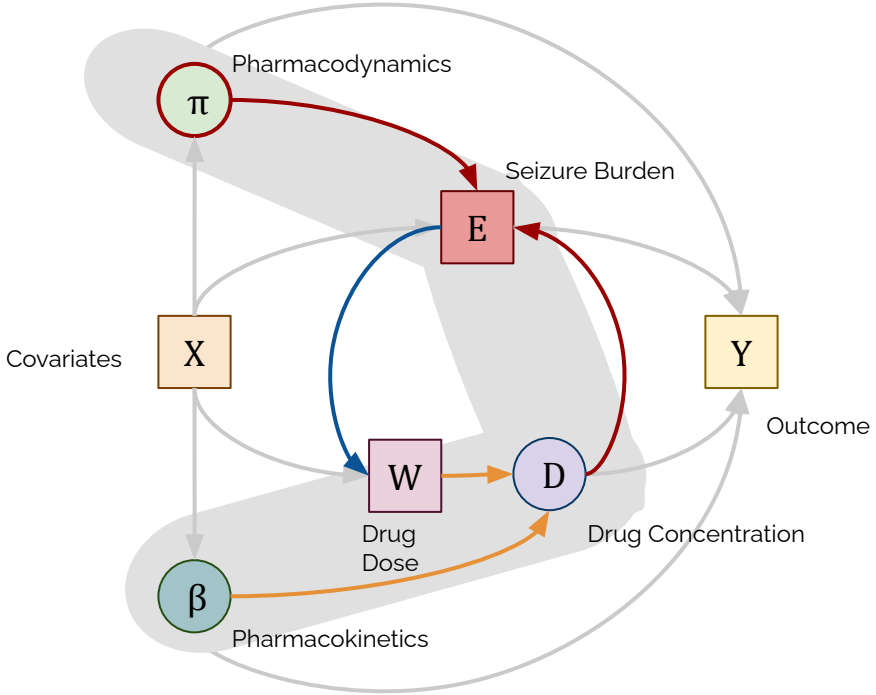
Mechanistic Modeling of Pharmacology

$$E_{i,t} = 1 - \sum_j \frac{D_{i,t,j}^{\alpha_{i,j}}}{D_{i,t,j}^{\alpha_{i,j}} + ED_{50,i,j}^{\alpha_{i,j}}}$$

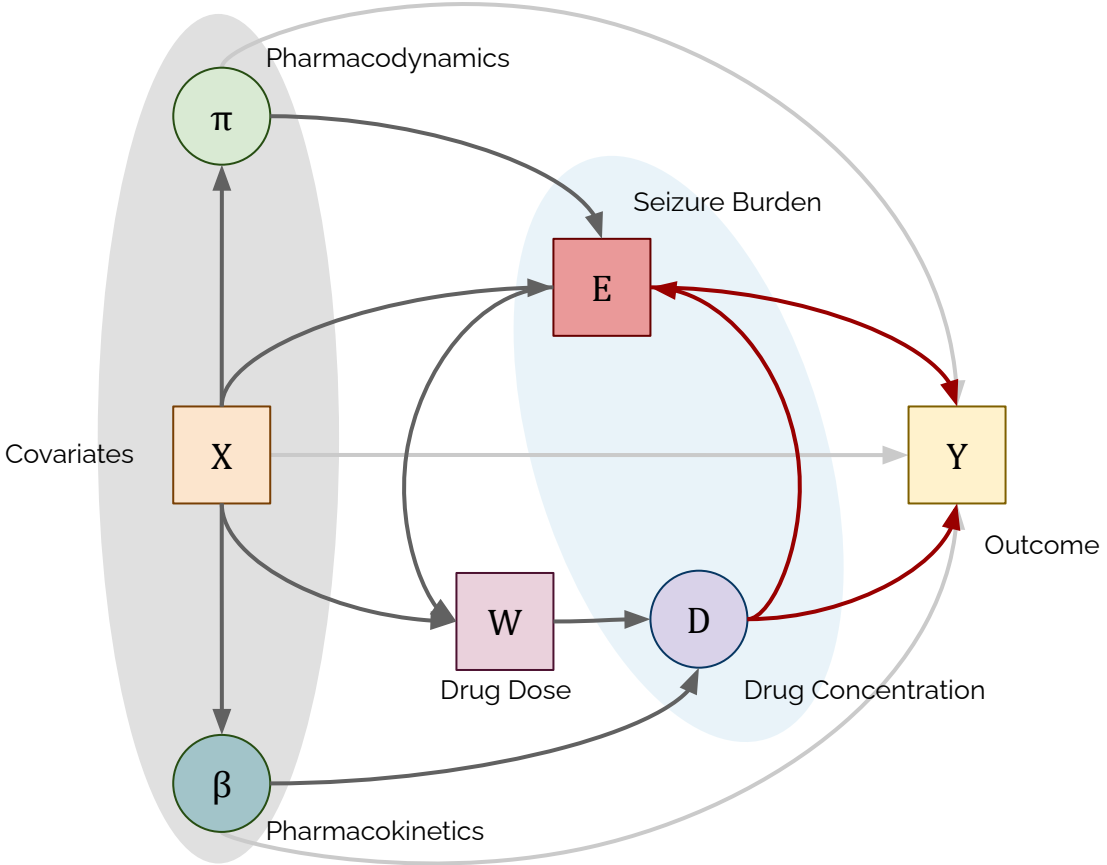
$$\pi_i = \{ED_{50,i,j}; \alpha_{i,j}\}_j$$

$$\frac{dD_{i,t,j}}{dt} = -\frac{1}{\beta_j} D_{i,t,j} + W_{i,t,j}$$

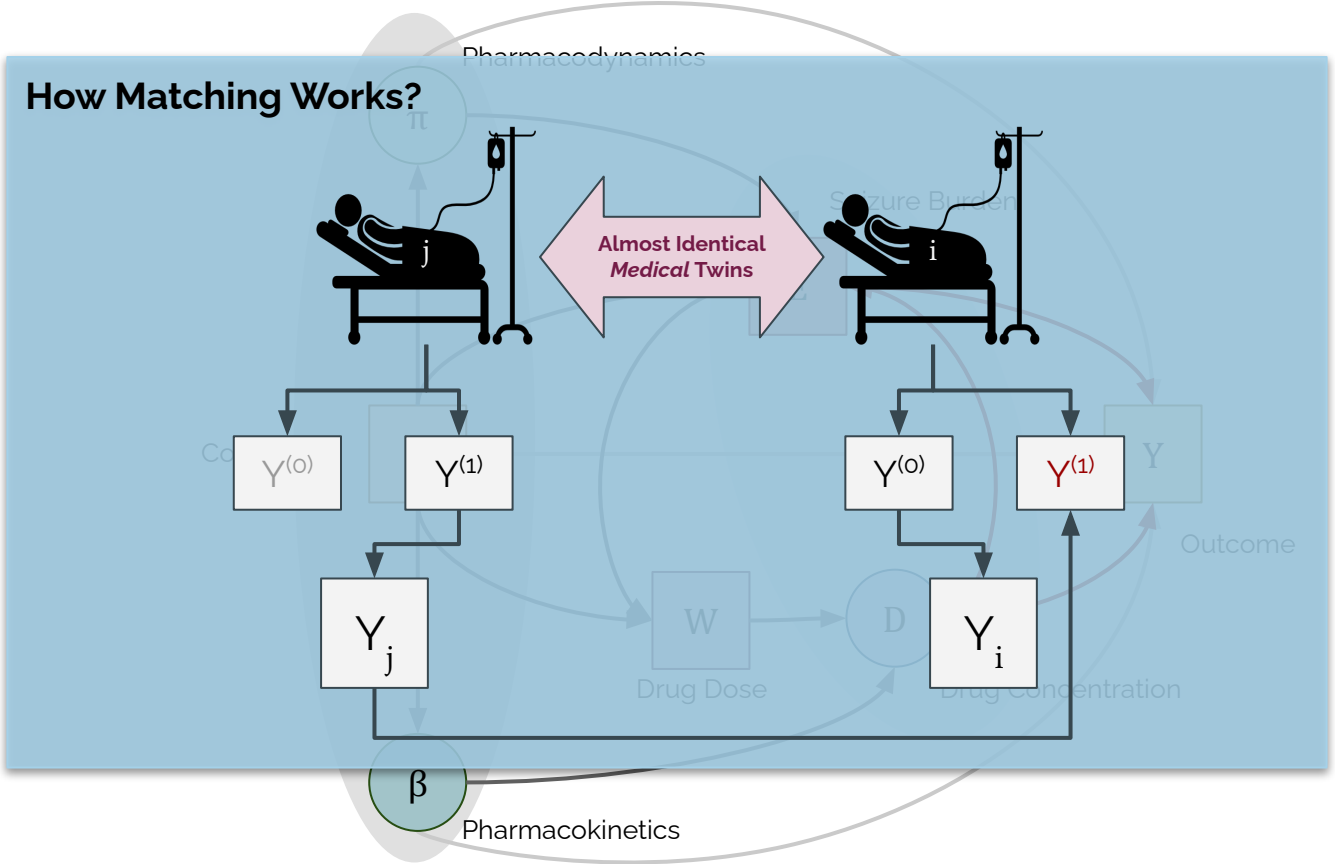
Drug Discharge Rate



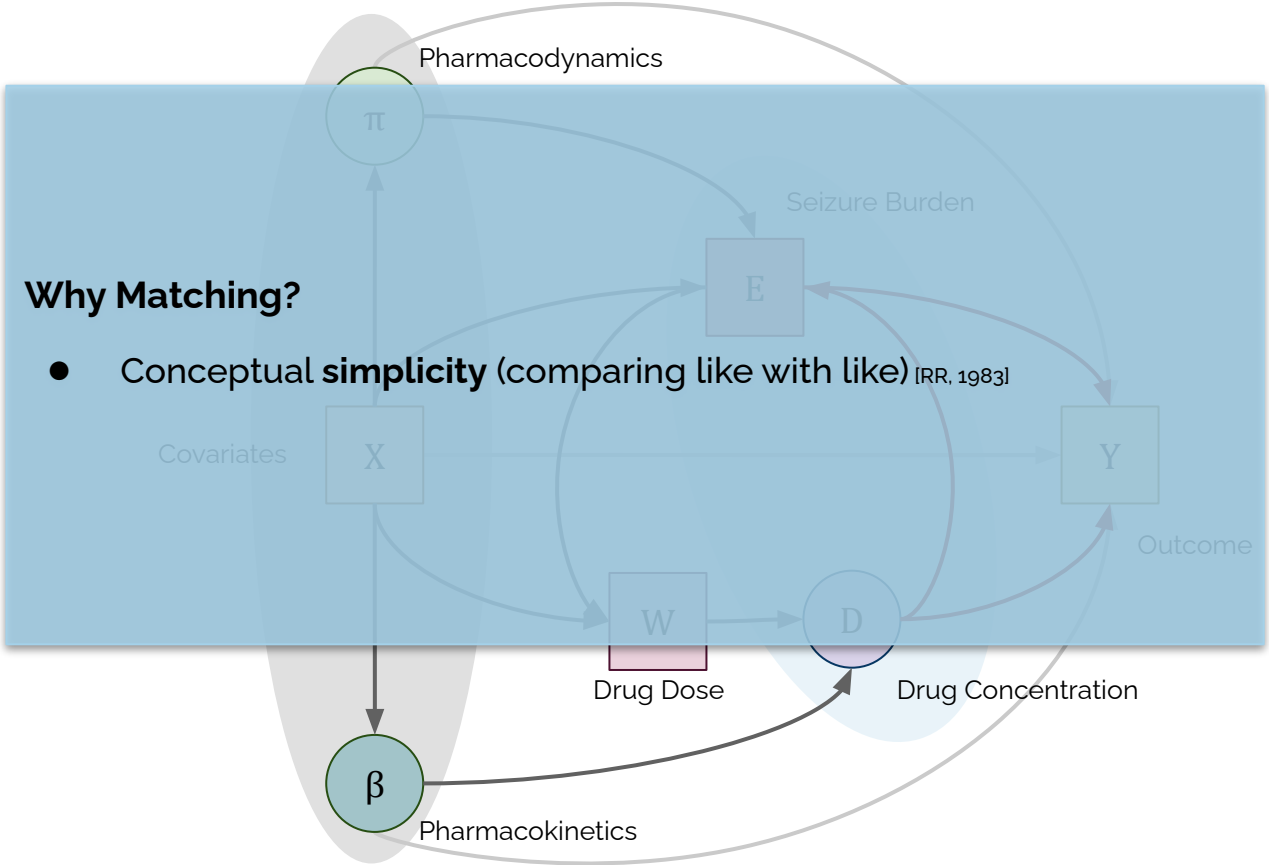
Controlling for Confounders: Covariates & Pharmacology



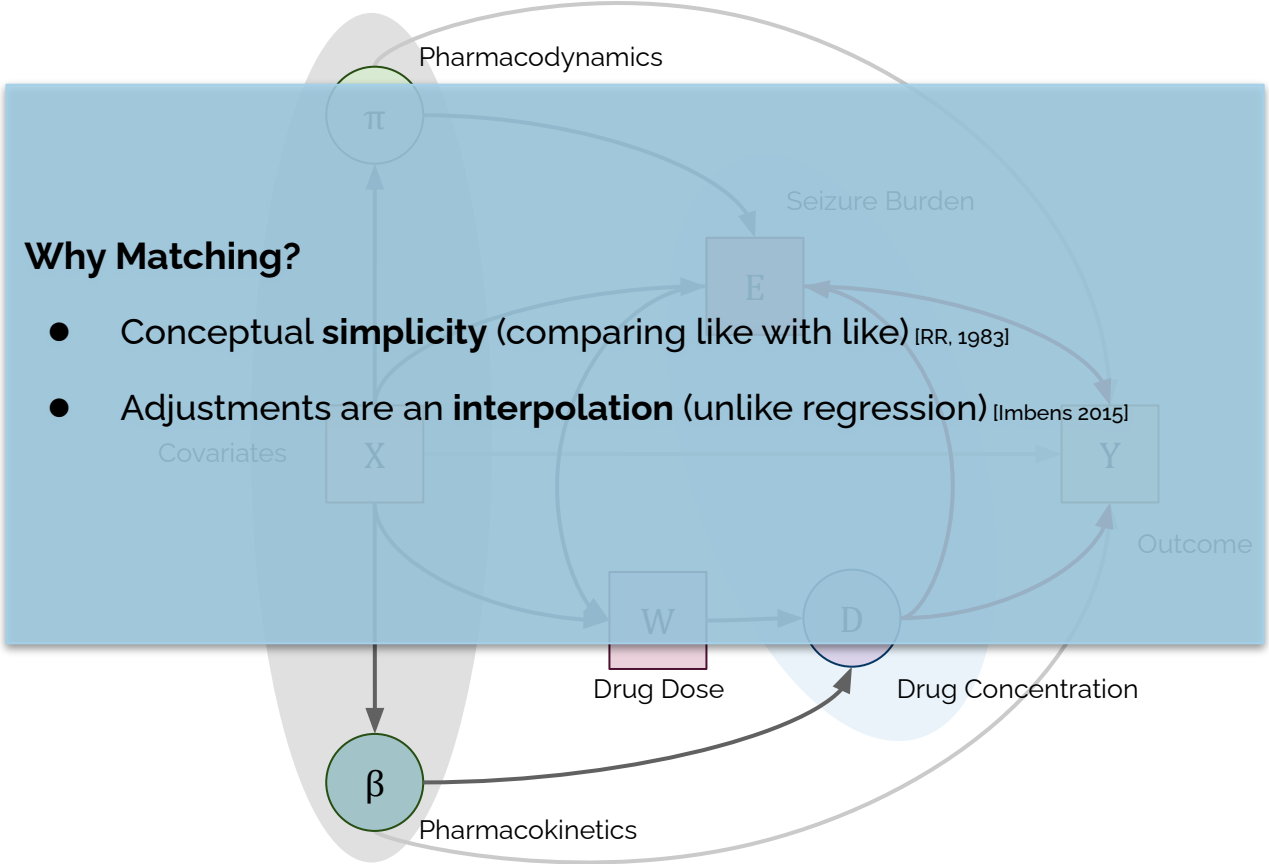
Controlling for Confounders: Covariates & Pharmacology



Controlling for Confounders: Covariates & Pharmacology



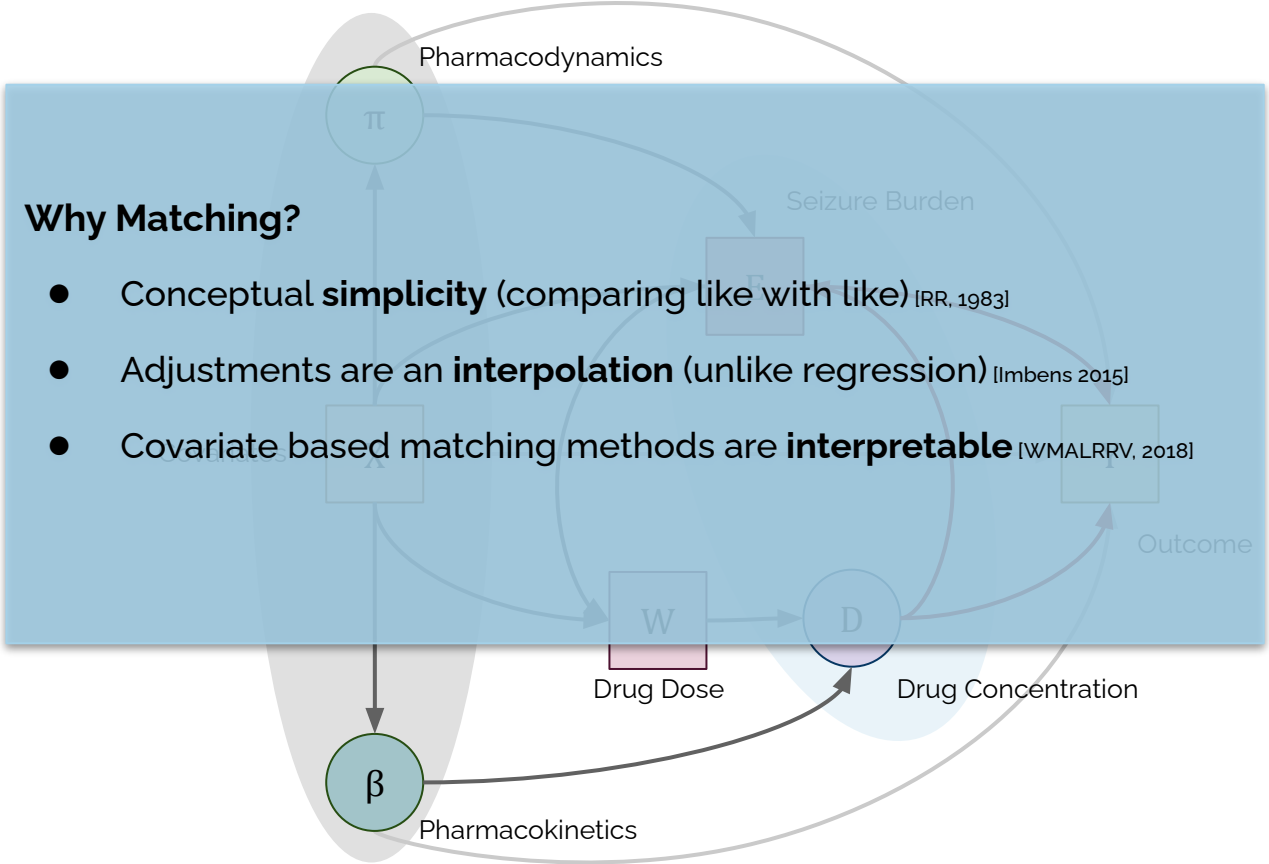
Controlling for Confounders: Covariates & Pharmacology



Why Matching?

- Conceptual **simplicity** (comparing like with like) [RR, 1983]
- Adjustments are an **interpolation** (unlike regression) [Imbens 2015]

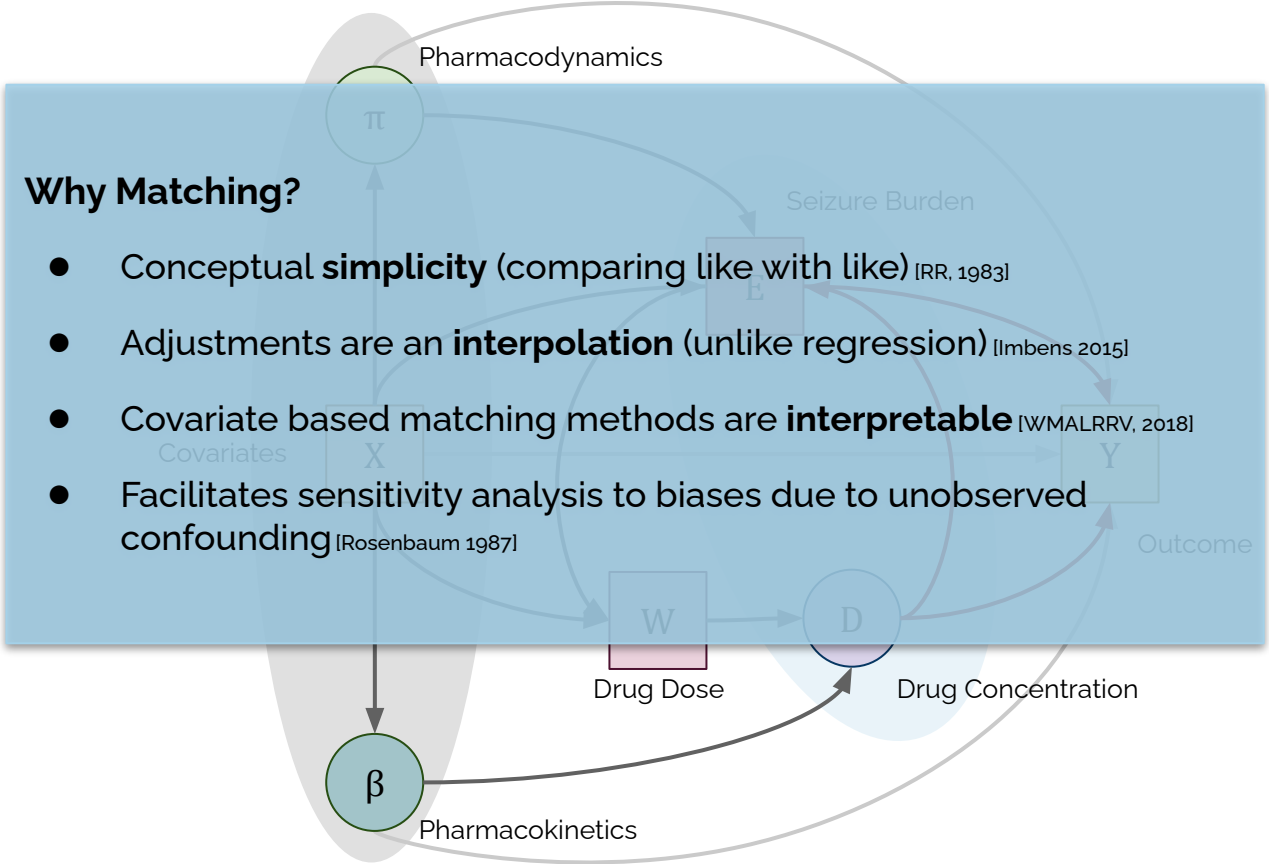
Controlling for Confounders: Covariates & Pharmacology



Why Matching?

- Conceptual **simplicity** (comparing like with like) [RR, 1983]
- Adjustments are an **interpolation** (unlike regression) [Imbens 2015]
- Covariate based matching methods are **interpretable** [WMALRRV, 2018]

Controlling for Confounders: Covariates & Pharmacology



Why Matching?

- Conceptual **simplicity** (comparing like with like) [RR, 1983]
- Adjustments are an **interpolation** (unlike regression) [Imbens 2015]
- Covariate based matching methods are **interpretable** [WMALRRV, 2018]
- Facilitates sensitivity analysis to biases due to unobserved confounding [Rosenbaum 1987]

Challenges with Matching: Toenail Problem

Name	ASM	Death (Y)	Hx Epilepsy (X1)	Toenail-Length (X2)
Jim	1	0	1	0.2 inch
Joe	0	0	0	0.1 inch
Jill	0	1	1	1 inch

Challenges with Matching: Toenail Problem

Name	ASM	Death (Y)	Hx Epilepsy (X1)	Toenail-Length (X2)
Jim	1	0	1	**0.2 inch**
Joe	0	0	0	**0.1 inch**
Jill	0	1	1	1 inch

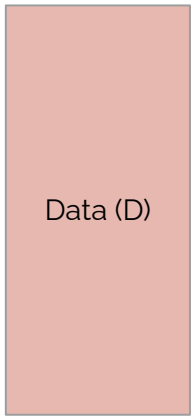
Challenges with Matching: Toenail Problem

Name	ASM	Death (Y)	Hx Epilepsy (X1)	Toenail-Length (X2)
Jim	1	0	**1**	0.2 inch
Joe	0	0	0	0.1 inch
Jill	0	1	**1**	1 inch

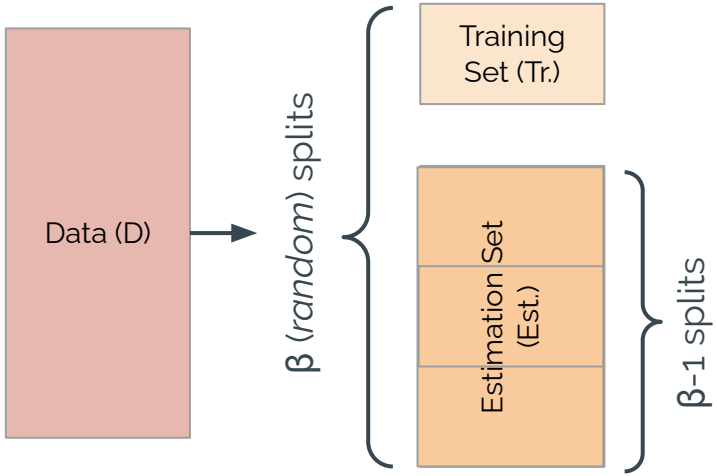
MALTS: Matching After Learning to Stretch
(Interpretable-and-Accurate Estimation Framework)

H Parikh, C Rudin, A Volfovsky - Journal of Machine Learning Research, 2022

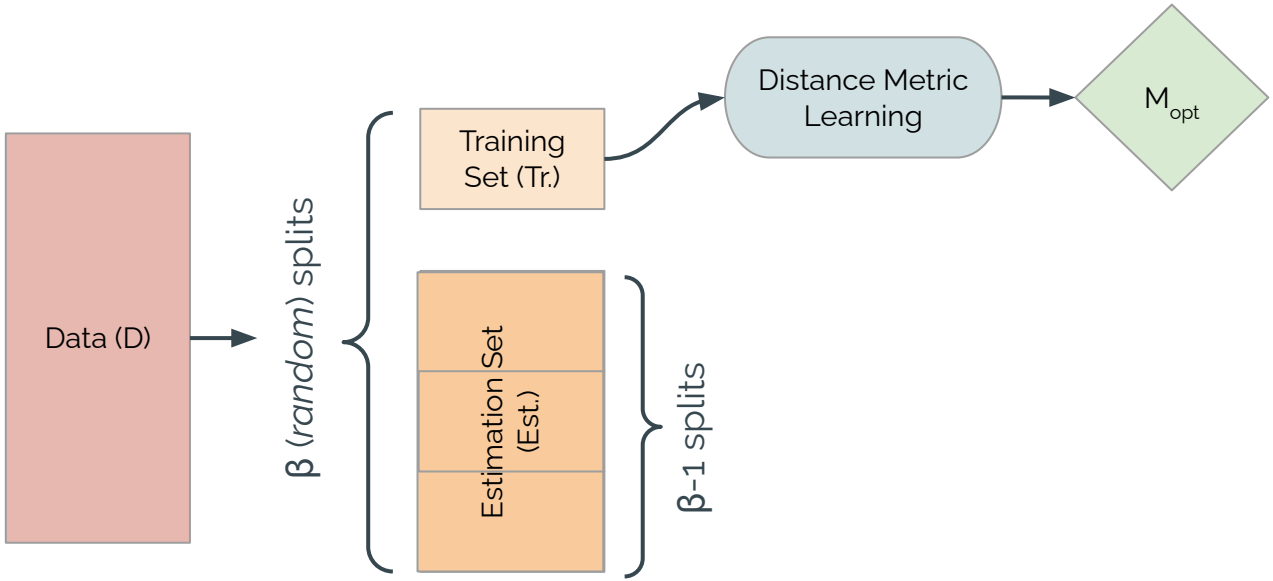
MALTS Framework



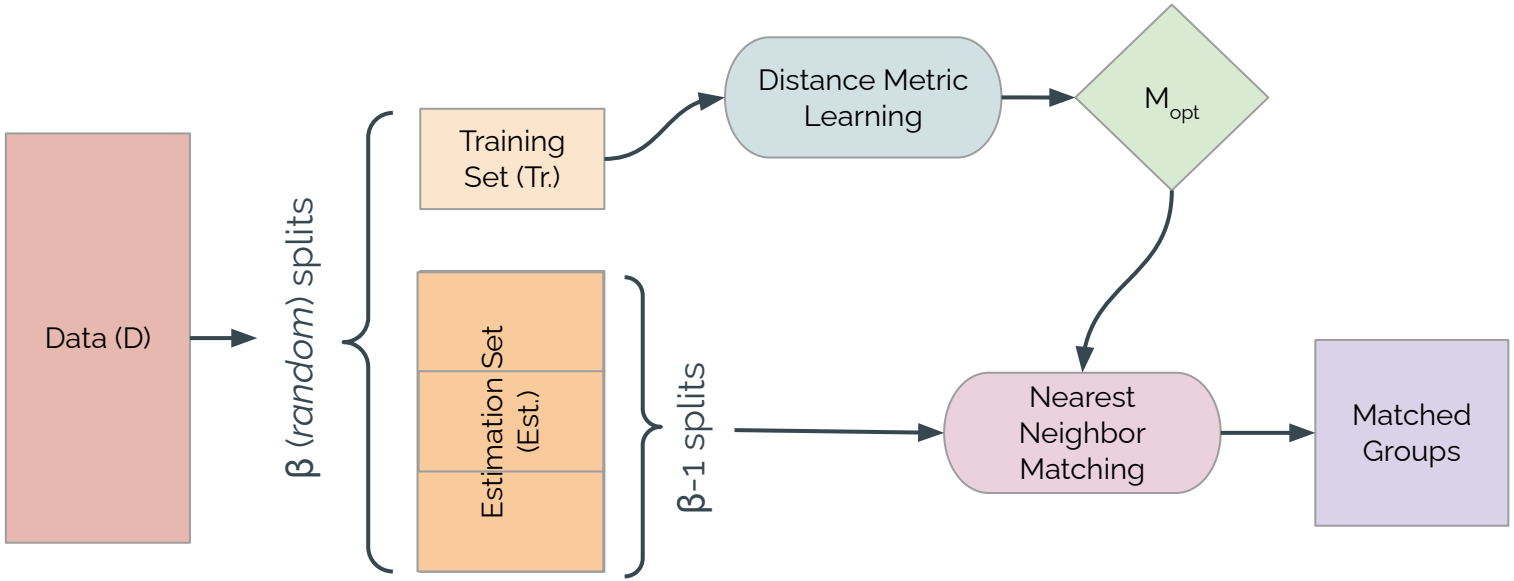
MALTS Framework



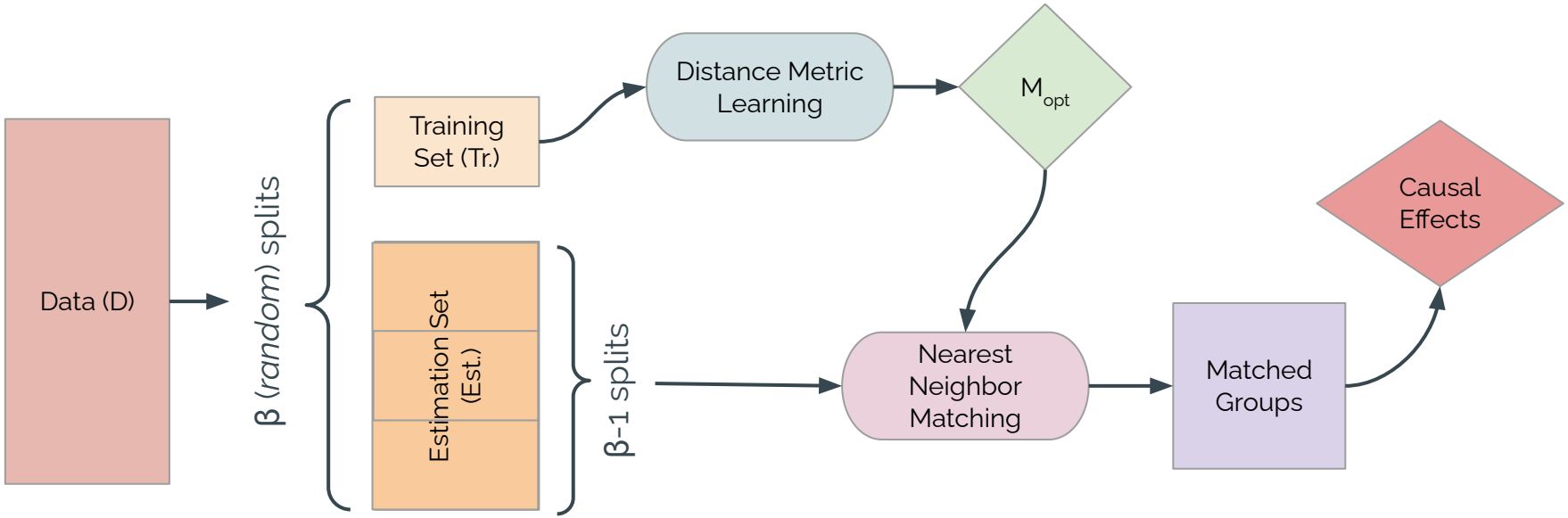
MALTS Framework



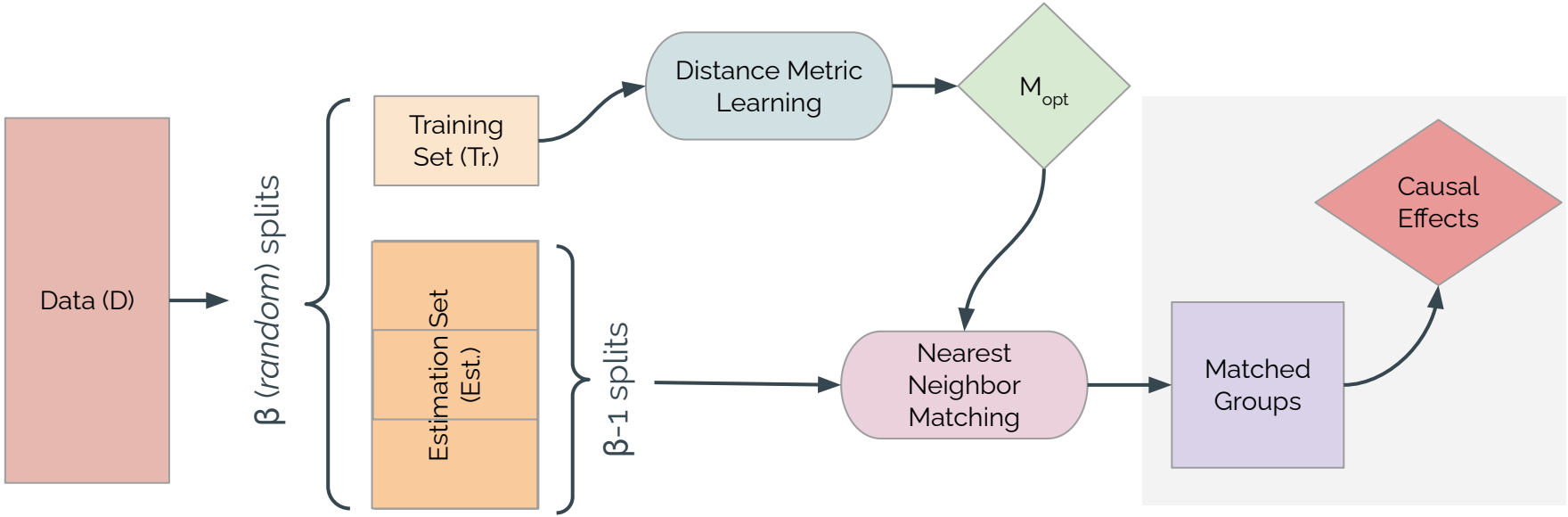
MALTS Framework



MALTS Framework



Estimation


Data (D)
β (random) splits
Training Set (Tr.)
Estimation Set (Est.)
β-1 splits
Distance Metric Learning
M_{opt}
Nearest Neighbor Matching
Matched Groups
Causal Effects

Estimation

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right)\right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation

Query Unit

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right)\right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation

Query Unit Distance Metric

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right)\right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation

Query Unit

Distance Metric

Set of Units with treatment t'

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right)\right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation

Query Unit

Distance Metric

Set of Units with treatment t'

Number of nearest neighbors

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right)\right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation

Query Unit

Distance Metric

Set of Units with treatment t'

Number of nearest neighbors

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

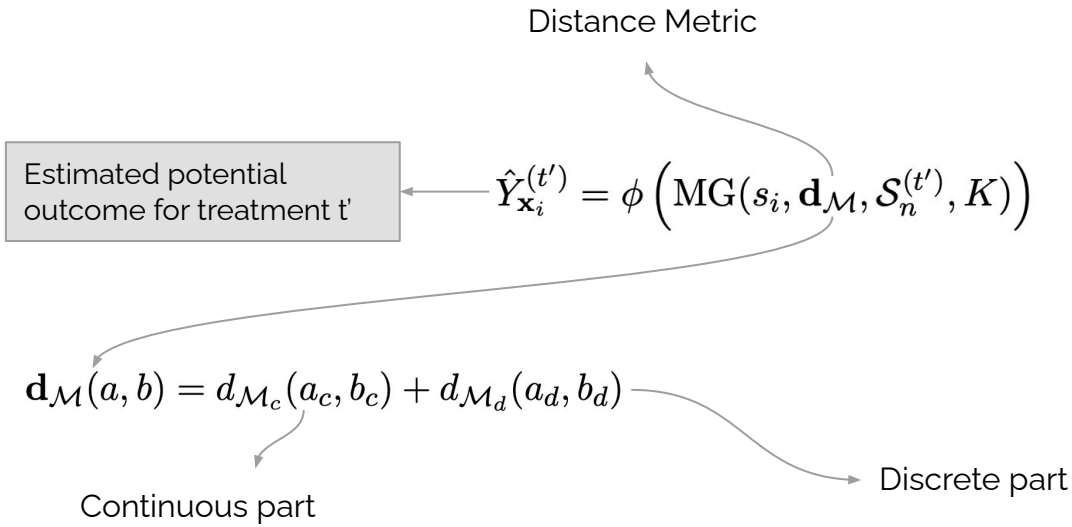
Estimator *e.g. mean, linear regression etc.*

where,

$$\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K) = KNN_{\mathcal{M}}^{\mathcal{S}_n}(\mathbf{x}_i, t') := \left\{ s_k : \left[\sum_{s_l \in \mathcal{S}_n^{(t')}} \mathbb{1}\left(\mathbf{d}_{\mathcal{M}}(\mathbf{x}_l, \mathbf{x}_i) < \mathbf{d}_{\mathcal{M}}(\mathbf{x}_k, \mathbf{x}_i)\right) \right] < K \right\}$$

K Nearest Neighbors with treatment t'

Estimation


Distance Metric
Estimated potential outcome for treatment t'
$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$
$\mathbf{d}_{\mathcal{M}}(a,b) = d_{\mathcal{M}_c}(a_c,b_c) + d_{\mathcal{M}_d}(a_d,b_d)$
Continuous part
Discrete part

Estimation

Estimated potential outcome for treatment t'

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

$$\mathbf{d}_{\mathcal{M}}(a, b) = d_{\mathcal{M}_c}(a_c, b_c) + d_{\mathcal{M}_d}(a_d, b_d)$$

Continuous part

$$d_{\mathcal{M}_c}(a_c, b_c) = \|\mathcal{M}_c a_c - \mathcal{M}_c b_c\|_2$$

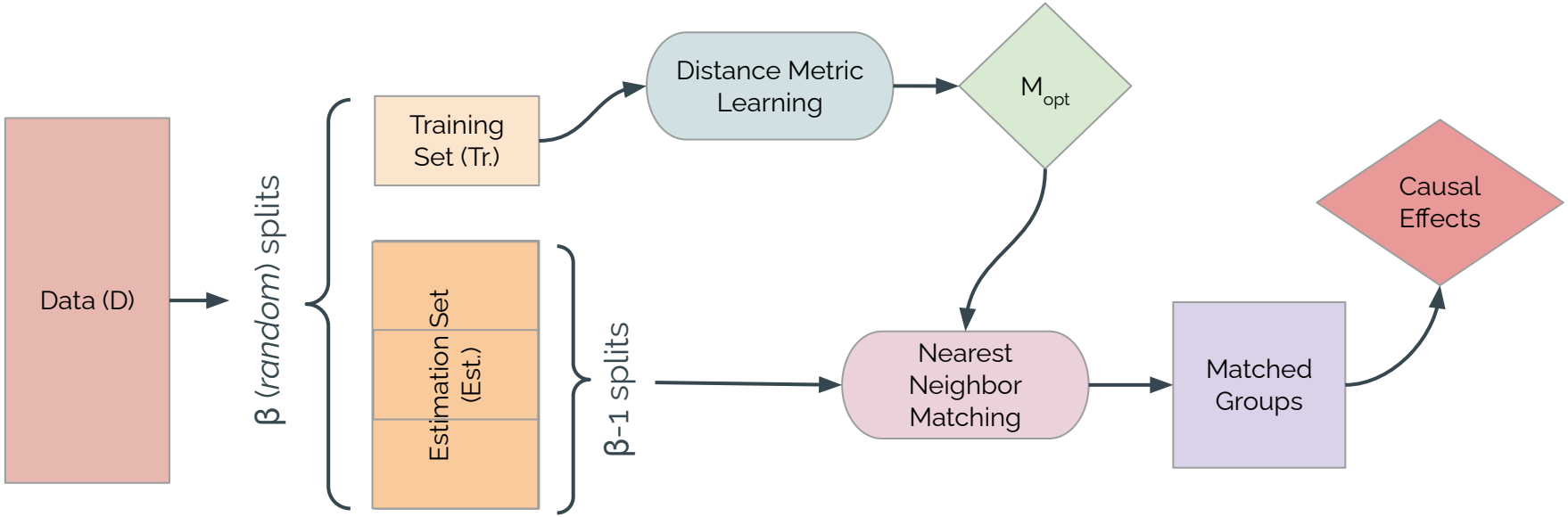
Stretched Euclidean Norm

Discrete part

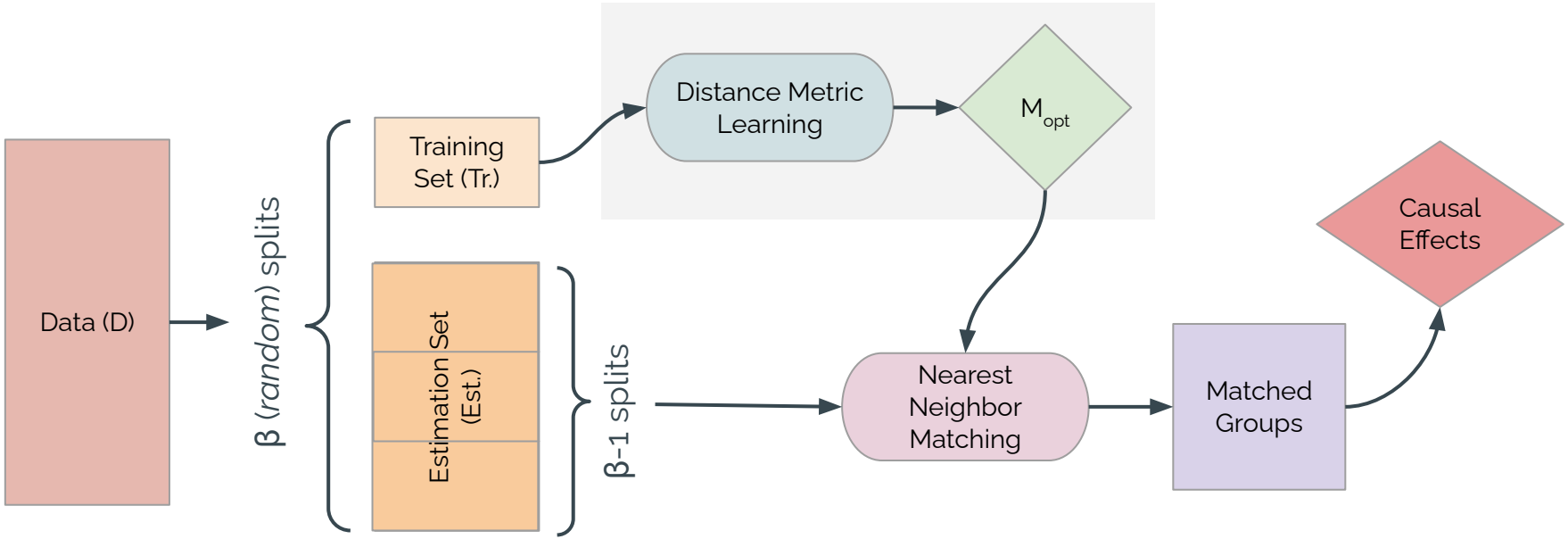
$$d_{\mathcal{M}_d}(a_d, b_d) = \sum_{j=0}^{|a_d|} \mathcal{M}_d^{(j,j)} \mathbb{1}[a_d^{(j)} \neq b_d^{(j)}]$$

Weighted Hamming Distance

Training


Data (D)
β (random) splits
Training Set (Tr.)
Estimation Set (Est.)
β-1 splits
Distance Metric Learning
M_{opt}
Nearest Neighbor Matching
Matched Groups
Causal Effects

Training


Data (D)
β (random) splits
Training Set (Tr.)
Estimation Set (Est.)
β-1 splits
Distance Metric Learning
M_{opt}
Nearest Neighbor Matching
Matched Groups
Causal Effects

Training

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

Find a distance metric d_M that *minimizes* the discrepancy between the *estimated* and the *true* treatment effect:

$$\min_{\mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(\underbrace{\widehat{Y}_{\mathbf{x}_i}^{(T)} - \widehat{Y}_{\mathbf{x}_i}^{(C)}}_{\textbf{Estimated Treatment Effect}}, \underbrace{Y_i^{(T)} - Y_i^{(C)}}_{\textbf{True Treatment Effect}}\Big)$$

Training

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

Find a distance metric d_M that *minimizes* the discrepancy between the *estimated* and the *true* treatment effect:

$$\min_{\mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\widehat{Y}_{\mathbf{x}_i}^{(T)} - \widehat{Y}_{\mathbf{x}_i}^{(C)}, Y_i^{(T)} - Y_i^{(C)}\right)$$

However, this loss *cannot* be evaluated because we observe only one of the potential outcomes.

Training

$$\hat{Y}_{\mathbf{x}_i}^{(t')} = \phi\left(\mathrm{MG}(s_i, \mathbf{d}_{\mathcal{M}}, \mathcal{S}_n^{(t')}, K)\right)$$

Find a distance metric d_M that *minimizes* the discrepancy between the *estimated* and the *true* treatment effect:

$$\min_{\mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \ell(\widehat{Y}_{\mathbf{x}_i}^{(T)} - \widehat{Y}_{\mathbf{x}_i}^{(C)}, Y_i^{(T)} - Y_i^{(C)})$$

However, this loss *cannot* be evaluated because we observe only one of the potential outcomes.

Surrogate Objective

$$\min_{\mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} t_i \ell(\widehat{Y}_{\mathbf{x}_i}^{(T)}, y_i) + (1 - t_i)\ell(\widehat{Y}_{\mathbf{x}_i}^{(C)}, y_i)$$

Training

$$\mathcal{M}(\mathcal{S}_{tr}) \in \arg\min_{\mathcal{M}} \left(c\|\mathcal{M}\|_{\mathcal{F}} + \Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) + \Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) \right)$$

$$\Delta^{(t)}_{\mathcal{S}_{tr}}(\mathcal{M}) := \frac{1}{|\mathcal{S}^{(t)}_{tr}|} \sum_{s_i \in \mathcal{S}^{(t)}_{tr}} \left| y_i - \sum_{s_l \in \mathcal{S}^{(t)}_{tr}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(t)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} y_l \right|$$

Training

Regularization loss

Loss for control units

Loss for treated units

$$\mathcal{M}(\mathcal{S}_{tr}) \in \arg\min_{\mathcal{M}} \left(c\|\mathcal{M}\|_{\mathcal{F}} + \Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) + \Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) \right)$$

$$\Delta^{(t)}_{\mathcal{S}_{tr}}(\mathcal{M}) := \frac{1}{|\mathcal{S}^{(t)}_{tr}|} \sum_{s_i \in \mathcal{S}^{(t)}_{tr}} \left| y_i - \sum_{s_l \in \mathcal{S}^{(t)}_{tr}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(t)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} y_l \right|$$

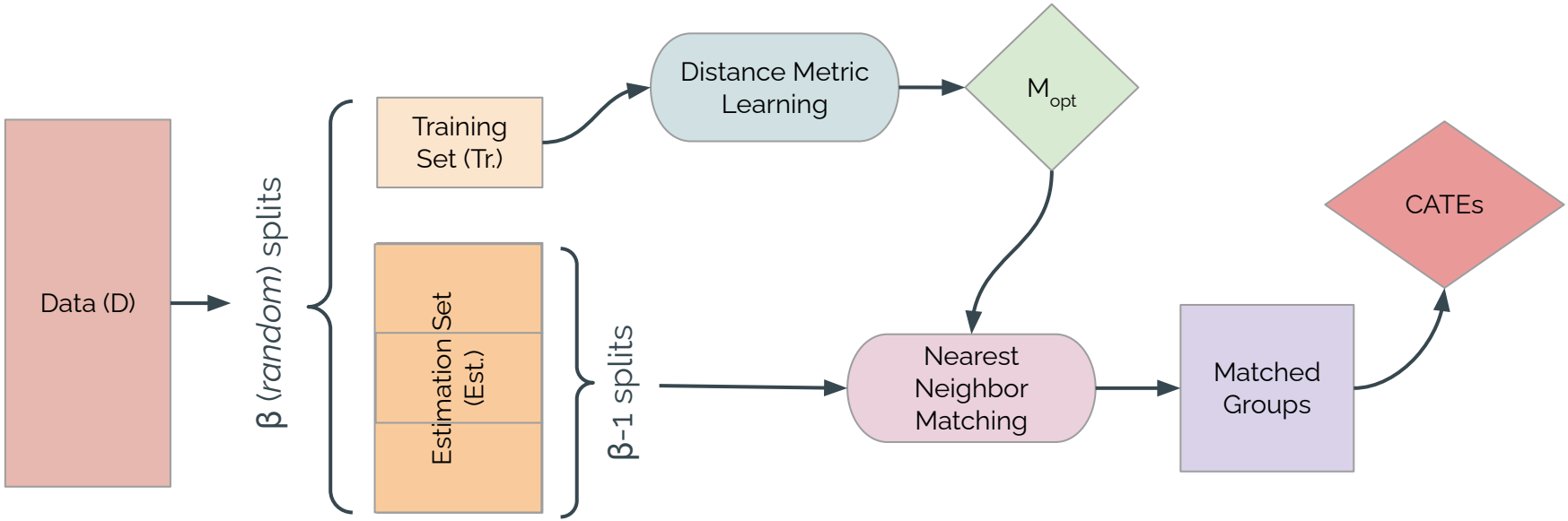
Training

$$\mathcal{M}(\mathcal{S}_{tr}) \in \arg\min_{\mathcal{M}} \left(c\|\mathcal{M}\|_{\mathcal{F}} + \Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) + \Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) \right)$$

Weighted Nearest Neighbors Estimate

$$\Delta^{(t)}_{\mathcal{S}_{tr}}(\mathcal{M}) := \frac{1}{|\mathcal{S}^{(t)}_{tr}|} \sum_{s_i \in \mathcal{S}^{(t)}_{tr}} \underbrace{\left| y_i - \sum_{s_l \in \mathcal{S}^{(t)}_{tr}} \frac{e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)}}{\sum_{s_k \in \mathcal{S}^{(t)}_{tr}} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_k)}} y_l \right|}_{\ell(\widehat{Y}^{(t)}_{\mathbf{x}_i}, y_i)}$$

Review: MALTS Framework



Experiments

Experiment 1: MALTS has **accuracy** on par with black box models

- Continuous and discrete covariates
- Highly non-linear process

Experiment 2: MALTS is end-to-end **interpretable**

- Compare matched groups across different matching methods

Experiments

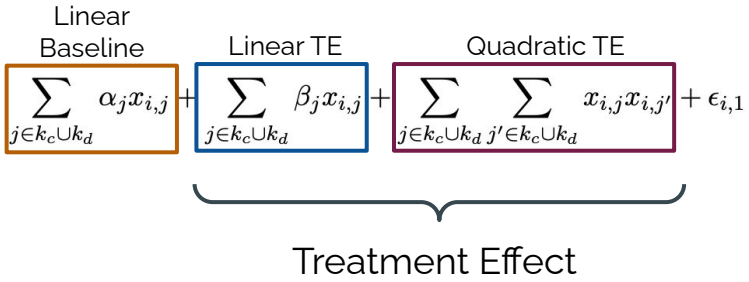
Experiment 1: MALTS has accuracy on-par with black box models

- **Continuous and discrete covariates**
- **Highly non-linear process**

Experiment 2: MALTS is end-to-end **interpretable**

- Compare matched groups across different matching methods

Quadratic Data Generation Process



Quadratic Data Generation Process

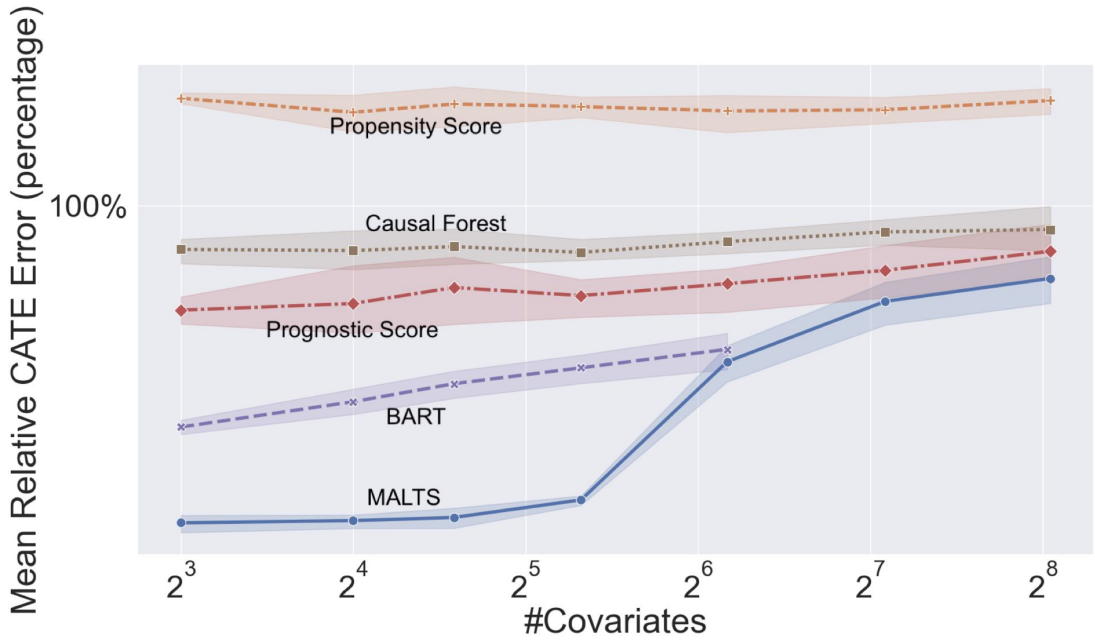
Linear Baseline

Linear TE

Quadratic TE

$$\sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \sum_{j \in k_c \cup k_d} \beta_j x_{i,j} + \sum_{j \in k_c \cup k_d} \sum_{j' \in k_c \cup k_d} x_{i,j} x_{i,j'} + \epsilon_{i,1}$$

MALTS is consistently accurate as the *dimensionality* increases.



Quadratic Data Generation Process

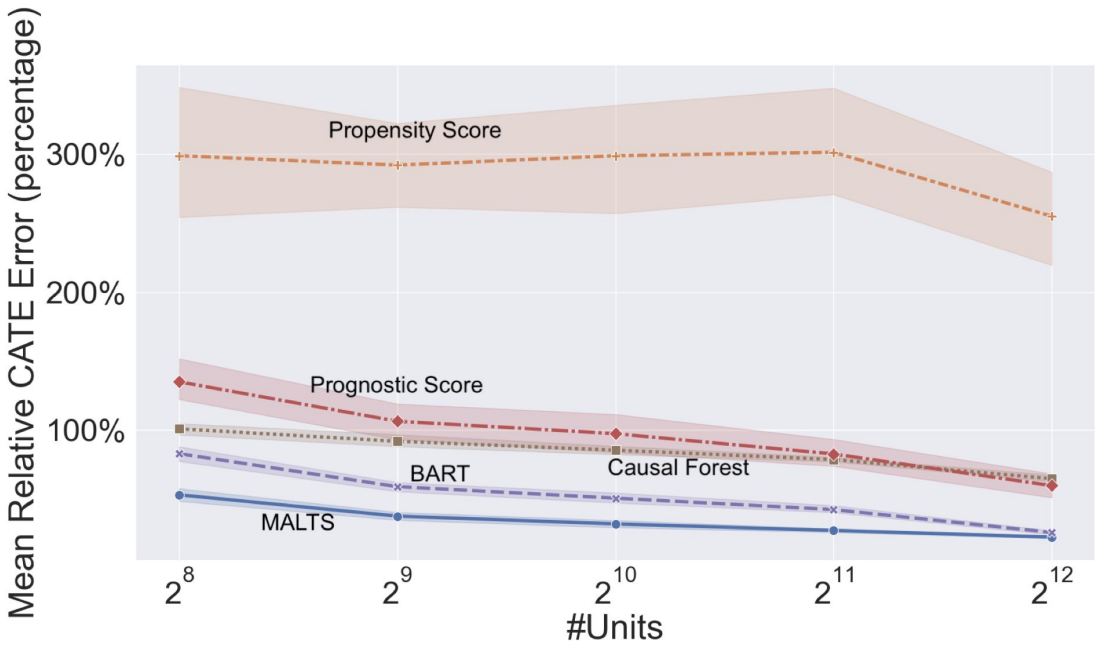
Linear Baseline

Linear TE

Quadratic TE

$$\sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \sum_{j \in k_c \cup k_d} \beta_j x_{i,j} + \sum_{j \in k_c \cup k_d} \sum_{j' \in k_c \cup k_d} x_{i,j} x_{i,j'} + \epsilon_{i,1}$$

MALTS is consistently accurate as the *sample size* increases.



Friedman's Data Generation Process

Baseline

Treatment Effect

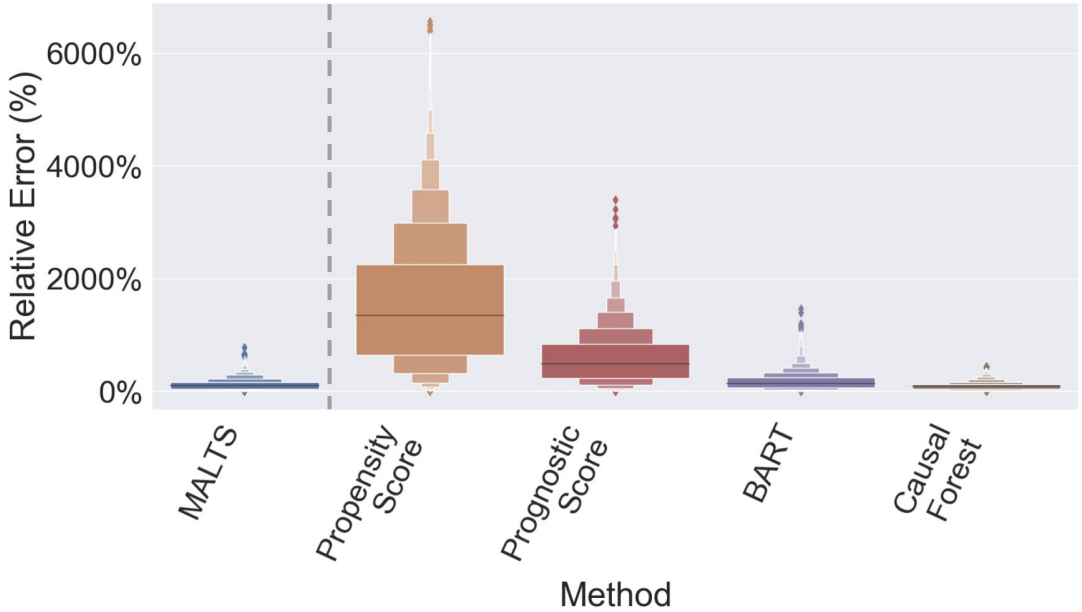
$$10\ \sin(\pi x_{i,1} x_{i,2}) + 20\ (x_{i,3} - 0.5)^2 + 10\ x_{i,4} + 5\ x_{i,5} + x_{i,3}\ \cos(\pi x_{i,1} x_{i,2}) + \epsilon_{i,1}$$

Friedman's Data Generation Process

Baseline

Treatment Effect

$$10\ \sin(\pi x_{i,1}x_{i,2}) + 20\ (x_{i,3} - 0.5)^2 + 10\ x_{i,4} + 5\ x_{i,5} + x_{i,3}\ \cos(\pi x_{i,1}x_{i,2}) + \epsilon_{i,1}$$



MALTS is on-par with black-box approaches even for highly non-linear data generation process.

Experiments

Experiment 1: MALTS has **accuracy** on par with black box models

- Continuous and discrete covariates
- Highly non-linear process

Experiment 2: MALTS is end-to-end interpretable

- **Compare matched groups across different matching methods**

Lalonde Experimental & Observational Study

Method	ATE Estimate
Truth	886

Lalonde Experimental & Observational Study

Method	ATE Estimate
Truth	886
GenMatch	549.53
Propensity Score	513.79
Prognostic Score	-897.76
BART-CV	713.20
Causal Forest-CV	-179.98

Lalonde Experimental & Observational Study

Method	ATE Estimate
Truth	886
GenMatch	549.53
Propensity Score	513.79
Prognostic Score	-897.76
BART-CV	713.20
Causal Forest-CV	-179.98
MALTS (pruned)	*891.75*

Lalonde Matched Groups

	Treatment	Covariates							Outcome
Unit ID	**Treated**	**Age**	**Education**	**Black**	**Hispanic**	**Married**	**No-Degree**	**Income-1975**	**Income-1978**
Query: 1	Yes	22	9	No	Yes	No	Yes	$0	$3596

Lalonde Matched Groups

	Treatment	Covariates							Outcome
Unit ID	**Treated**	**Age**	**Education**	**Black**	**Hispanic**	**Married**	**No-Degree**	**Income-1975**	**Income-1978**
Query: 1	Yes	22	9	No	Yes	No	Yes	$0	$3596
Prognostic Scores									
338	No	**44**	9	**Yes**	**No**	No	Yes	$0	$9722
340	No	22	**12**	**Yes**	**No**	No	**No**	**$532**	$1333
355	No	18	10	No	Yes	No	Yes	$0	$1859
Propensity Scores									
451	No	22	8	**Yes**	**No**	No	Yes	$0	$1391
330	No	22	8	No	Yes	No	Yes	$0	$9921
407	No	20	**12**	**Yes**	**No**	No	**No**	**$1371**	$20893

Lalonde Matched Groups

	Treatment	Covariates							Outcome
Unit ID	**Treated**	**Age**	**Education**	**Black**	**Hispanic**	**Married**	**No-Degree**	**Income-1975**	**Income-1978**
Query: 1	Yes	22	9	No	Yes	No	Yes	$0	$3596
Prognostic Scores									
338	No	**44**	9	**Yes**	**No**	No	Yes	$0	$9722
340	No	22	**12**	**Yes**	**No**	No	**No**	**$532**	$1333
355	No	18	10	No	Yes	No	Yes	$0	$1859
Propensity Scores									
451	No	22	8	**Yes**	**No**	No	Yes	$0	$1391
330	No	22	8	No	Yes	No	Yes	$0	$9921
407	No	20	**12**	**Yes**	**No**	No	**No**	**$1371**	$20893
Our Approach (MALTS)									
330	No	22	8	No	Yes	No	Yes	$0	$9921
299	No	22	9	**Yes**	**No**	No	Yes	$0	$0
416	No	22	9	**Yes**	**No**	No	Yes	$0	$12898

Theoretical Guarantees

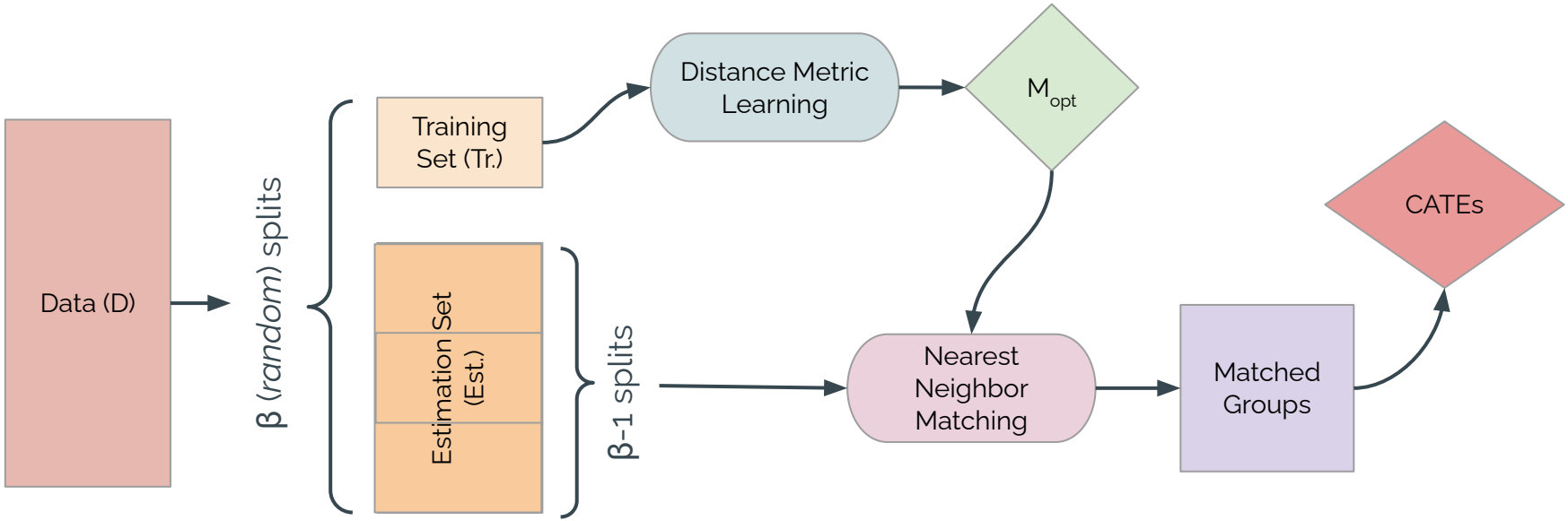
Theorem 1. (Consistency) Given a smooth distance metric, a K-nearest neighbors estimate of treatment effect *estimates* are accurate.

- Estimated treatment effect is consistent with the true treatment effect

Theorem 2. (Generalizability) MALTS *learned* distance metric is generalizable

- For the *learned distance metric*, the empirical loss on the training set is not far from population loss (with high probability)

Review: MALTS Framework



Back to

Seizures in ICU Patients

Understanding the Matched Groups

Learned Distance Metric

Learned Distance Metric

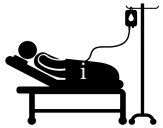
Pharmacodynamic parameters for an important ASM

Learned Distance Metric

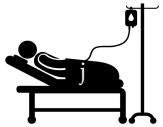
Extent of impaired consciousness
(in acute medical and trauma patients)

Measure of ICU mortality

Auditability: It's a Tight Match!



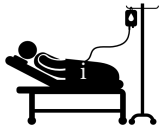
Age / Gender	59 / F
APACHE-II (Prognosis)	8
Major Medical Hx	Subarachnoid Hemorrhage
Rx Hill's Response (Levetiracetam)	1.26
…	…
Doctor's Notes	*Traumatic brain Injury. Prior witnessed episodes of seizure; admitted after an episode of motionless stare*



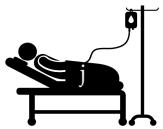
Age / Gender	62 / F
APACHE-II (Prognosis)	7
Major Medical Hx	Subarachnoid Hemorrhage
Rx Hill's Response (Levetiracetam)	1.23
…	…
Doctor's Notes	*Complicated neurological history; diagnosed with aseptic meningitis. Admitted after she was found unresponsive.*

Comparing Patients within a Matched Group

Auditability: More than just Numbers!



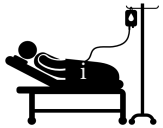
Age / **Gender**	64 / **F**
APACHE-II (Prognosis)	**3**
Major Medical Hx	Brain Tumor
Rx Hill's Response (Levetiracetam)	**1.22**
…	…
Doctor's Notes	*Brain tumor has grown larger and is causing swelling in the brain*



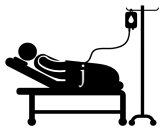
Age / **Gender**	28 / **F**
APACHE-II (Prognosis)	**3**
Major Medical Hx	Severe Pneumonia
Rx Hill's Response (Levetiracetam)	**1.20**
…	…
Doctor's Notes	*Large brain blood vessel malformation that has caused focal seizures. Admitted due to pain and bleeding in the right eye.*

Comparing Patients within a Matched Group

Auditability: More than just Numbers!



Age / **Gender**	64 / **F**
APACHE-II (Prognosis)	**3**
Major Medical Hx	Brain Tumor
Rx Hill's Response (Levetiracetam)	**1.22**
…	…
Doctor's Notes	*Brain tumor has grown larger and is causing swelling in the brain*



Age / **Gender**	28 / **F**
APACHE-II (Prognosis)	**3**
Major Medical Hx	Severe Pneumonia
Rx Hill's Response (Levetiracetam)	**1.20**
…	…
Doctor's Notes	*Large brain blood vessel malformation that has caused focal seizures. Admitted due to pain and bleeding in the right eye.*

Doctors' Assessment

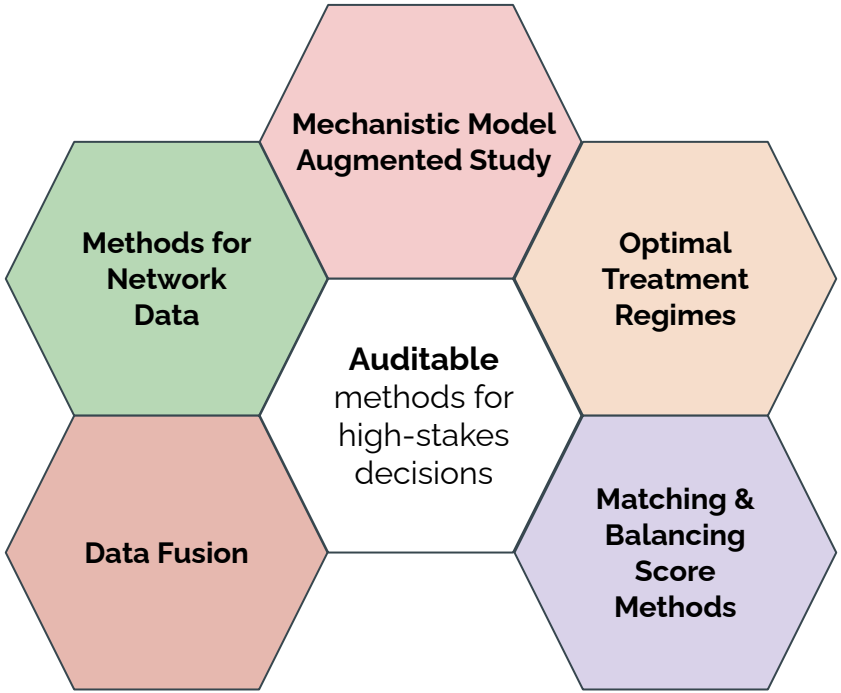
- One of the patients is much younger, but her history of **severe chronic illness makes her comparable** to the other patient.
- Both patients have **relatively high risk for seizures**.
- Based on data available at hospital admission, these patients with history of epilepsy or relatively static neurological injury have **good short term prognosis** compared.

Comparing Patients within a Matched Group

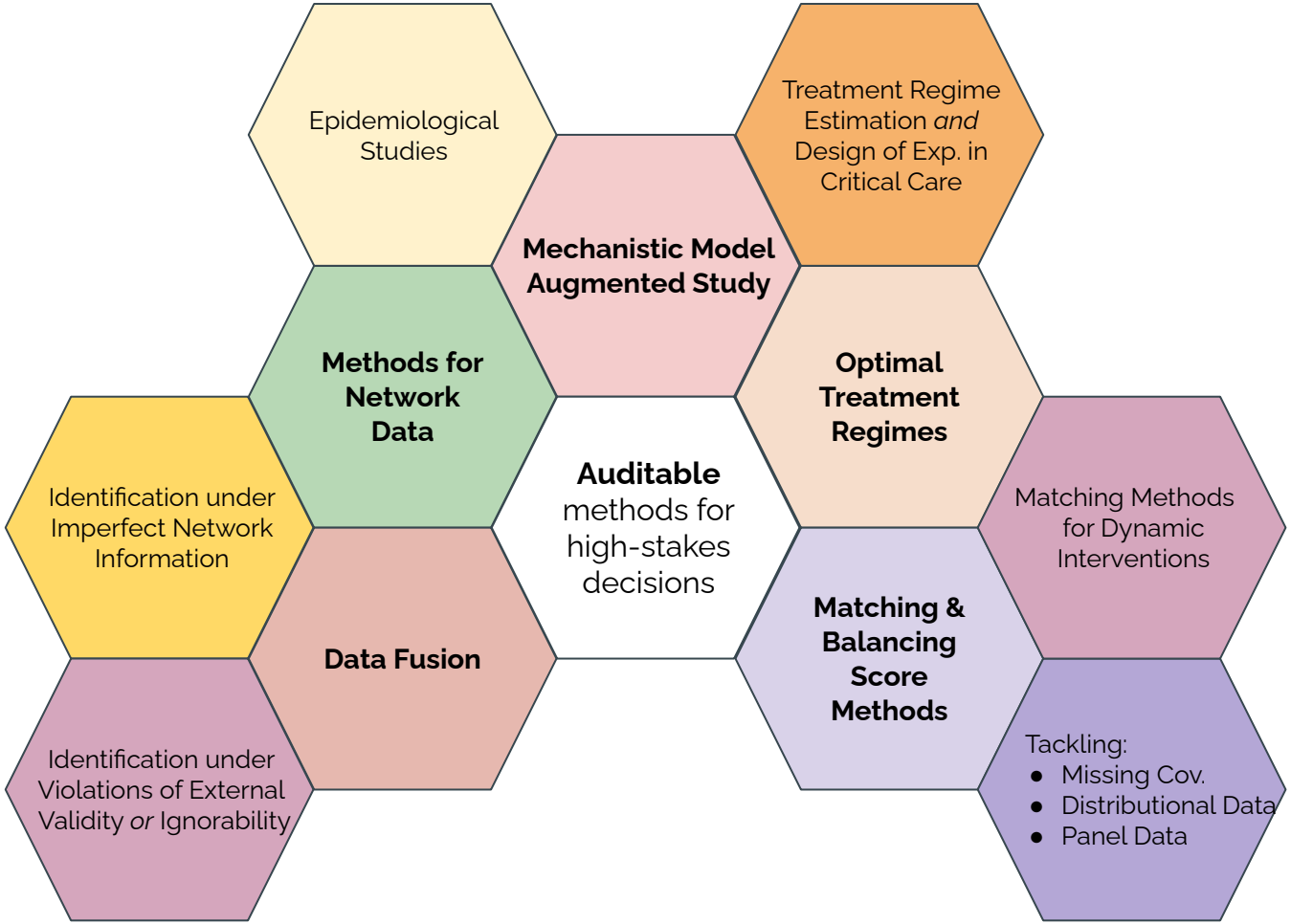
Qualitative Analysis

Future Directions

Future Directions



Future Directions



Advisors, Mentors, Collaborators and Friends!

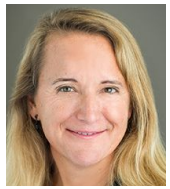

Cynthia Rudin


Alexander Volfovsky


Sudeepa Roy


Eric Tchetgen Tchetgen


M Brandon Westover


Lise Getoor


Dan Suciu


Babak Salimi


Marco Morucci


Kentaro Hoffman


Quinn Lanners


Amir Gilad


Vittorio Orlandi


Haoqi Sun


Sahar Zafar


Srikar Katta

Thank you!

Personal-website: https://sites.google.com/view/harshparikh/
Lab-website: https://almost-matching-exactly.github.io/
AME-Github: https://github.com/almost-matching-exactly/MALTS

More Slides!!!

MALTS Experiments

Experiments

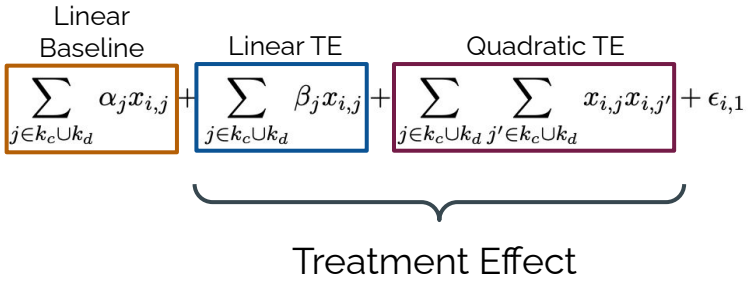
Experiment 1: MALTS has accuracy on-par with black box models

- **Continuous and discrete covariates**
- **Highly non-linear process**

Experiment 2: MALTS is end-to-end **interpretable**

- Compare matched groups across different matching methods

Quadratic Data Generation Process



Quadratic Data Generation Process

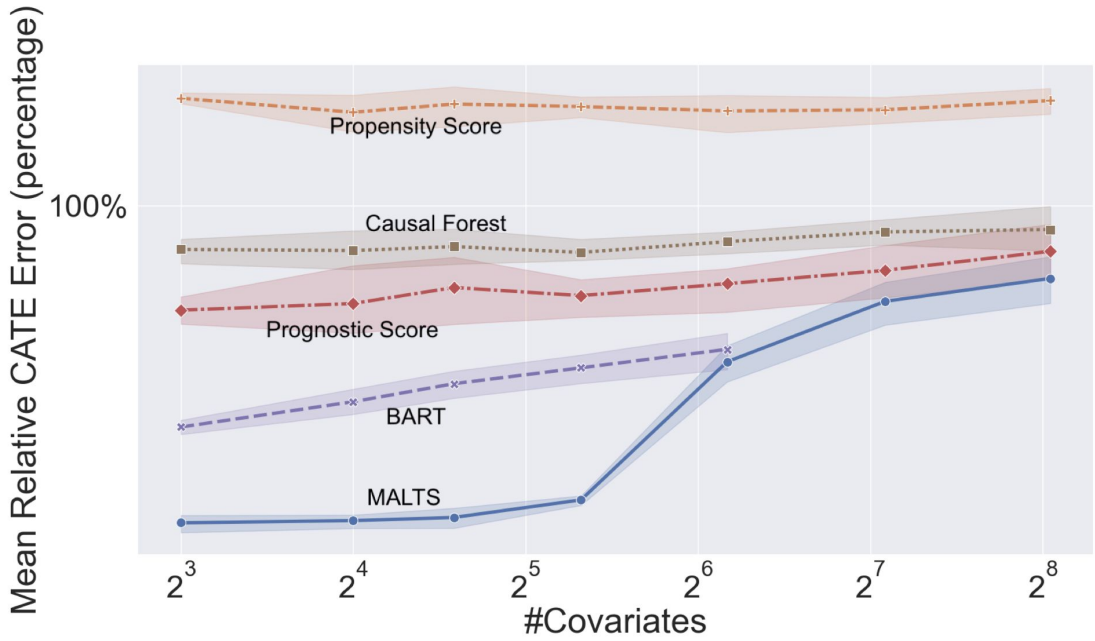
Linear Baseline

Linear TE

Quadratic TE

$$\sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \sum_{j \in k_c \cup k_d} \beta_j x_{i,j} + \sum_{j \in k_c \cup k_d} \sum_{j' \in k_c \cup k_d} x_{i,j} x_{i,j'} + \epsilon_{i,1}$$

MALTS is consistently accurate as the *dimensionality* increases.



Quadratic Data Generation Process

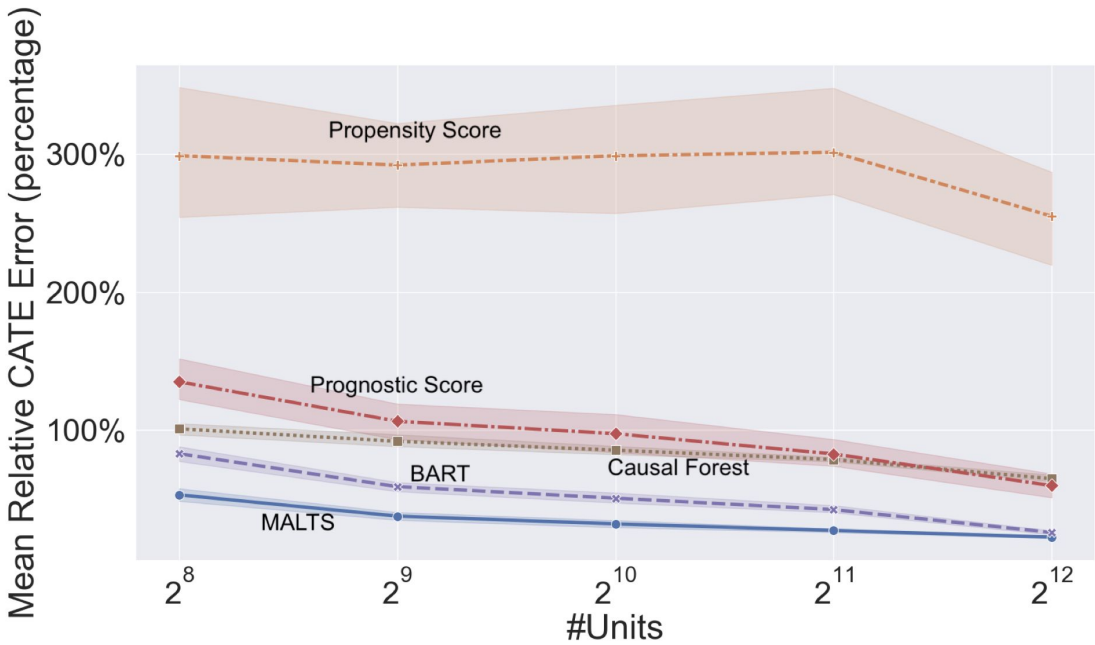
Linear Baseline

Linear TE

Quadratic TE

$$\sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \sum_{j \in k_c \cup k_d} \beta_j x_{i,j} + \sum_{j \in k_c \cup k_d} \sum_{j' \in k_c \cup k_d} x_{i,j} x_{i,j'} + \epsilon_{i,1}$$

MALTS is consistently accurate as the *sample size* increases.



Friedman's Data Generation Process

Baseline

Treatment Effect

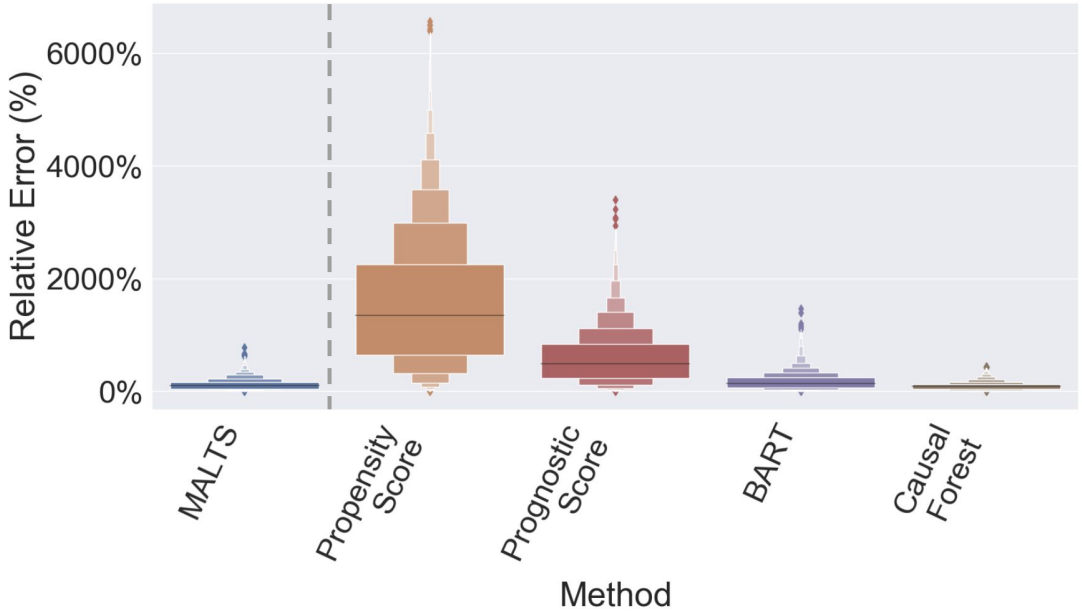
$$10\ \sin(\pi x_{i,1} x_{i,2}) + 20\ (x_{i,3} - 0.5)^2 + 10\ x_{i,4} + 5\ x_{i,5} + x_{i,3}\ \cos(\pi x_{i,1} x_{i,2}) + \epsilon_{i,1}$$

Friedman's Data Generation Process

Baseline

Treatment Effect

$$10\ \sin(\pi x_{i,1} x_{i,2}) + 20\ (x_{i,3} - 0.5)^2 + 10\ x_{i,4} + 5\ x_{i,5} + x_{i,3}\ \cos(\pi x_{i,1} x_{i,2}) + \epsilon_{i,1}$$



MALTS is on-par with black-box approaches even for highly non-linear data generation process.

MALTS Theory

Theoretical Guarantees

Theorem 1. Given a smooth distance metric, a K-nearest neighbors estimate of treatment effect *estimates* are accurate.

- **Estimated treatment effect is consistent with the true treatment effect**

Theorem 2. MALTS *learned* distance metric is generalizable

- For the *learned distance metric*, the empirical loss on the training set is not far from population loss (with high probability)

Smooth Distance Metric and Accurate CATEs

Smooth Distance Metric: If X_i and X_j are *close* under distance metric d_M, then **E**[$Y_i \mid X_i, T_i{=}t$] is also *close* to **E**[$Y_j \mid X_j, T_j{=}t$]

Smooth Distance Metric and Accurate CATEs

Smooth Distance Metric: If X_i and X_j are *close* under distance metric d_M, then **E**[$Y_i \mid X_i, T_i{=}t$] is also *close* to **E**[$Y_j \mid X_j, T_j{=}t$]

Given,

- Smooth distance metric d_M
- Covariate vector **x**
- Constant **α**

Theorem 1: If there exists a small enough caliper "a" and a large enough "N" then

$$P(|\widehat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \geq \alpha) \leq \delta_{d_{\mathcal{M}}}(\alpha, |MG(\mathbf{x}, a)|)$$

Smooth Distance Metric and Accurate CATEs

Smooth Distance Metric: If X_i and X_j are *close* under distance metric d_M, then **E**[$Y_i \mid X_i, T_i{=}t$] is also *close* to **E**[$Y_j \mid X_j, T_j{=}t$]

Given,

- Smooth distance metric d_M
- Covariate vector **x**
- Constant **α**

Theorem 1: If there exists a small enough caliper "a" and a large enough "N" then

$$P(|\widehat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \geq \alpha) \leq \delta_{d_{\mathcal{M}}}(\alpha, |MG(\mathbf{x}, a)|)$$

Proof Sketch: Use the smooth distance metric property with Hoeffding's to bound the potential outcome estimates and truth with high probability. Then use triangle inequality.

Theoretical Guarantees

Theorem 1. Given a smooth distance metric, a K-nearest neighbors estimate of treatment effect *estimates* are accurate.

- Estimated treatment effect is consistent with the true treatment effect

Theorem 2. MALTS *learned* distance metric is generalizable

- **For the *learned distance metric*, the empirical loss on the training set is not far from population loss (with high probability)**

Losses

Pairwise Loss

$$loss[\mathcal{M}, s_i, s_l] := \begin{cases} e^{-\mathbf{d}_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_l)} |y_i - y_l| & \text{if } t_i = t_l \\ \infty & \text{otherwise.} \end{cases}$$

Empirical Avg. Loss

$$L_{emp}(\mathcal{M}, \mathcal{S}_n) := \frac{1}{n^2} \sum_{(s_i, s_l) \in (\mathcal{S}_n \times \mathcal{S}_n)} loss[\mathcal{M}, s_i, s_l]$$

$$\Delta^{(C)}_{\mathcal{S}_{tr}}(\mathcal{M}) \leq \frac{1}{\exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right) \left|\mathcal{S}_{tr}^{(C)}\right|^2} \sum_{s_i \in \mathcal{S}_{tr}^{(C)}} \sum_{s_l \in \mathcal{S}_{tr}^{(C)}} loss[\mathcal{M}, s_i, s_l] = \frac{L_{emp}(\mathcal{M}, \mathcal{S}_{tr}^{(C)})}{\exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right)}$$

$$\Delta^{(T)}_{\mathcal{S}_{tr}}(\mathcal{M}) \leq \frac{L_{emp}(\mathcal{M}, \mathcal{S}_{tr}^{(T)})}{\exp\left(-\frac{g_0 \mathbf{C}_x^2}{c}\right)}$$

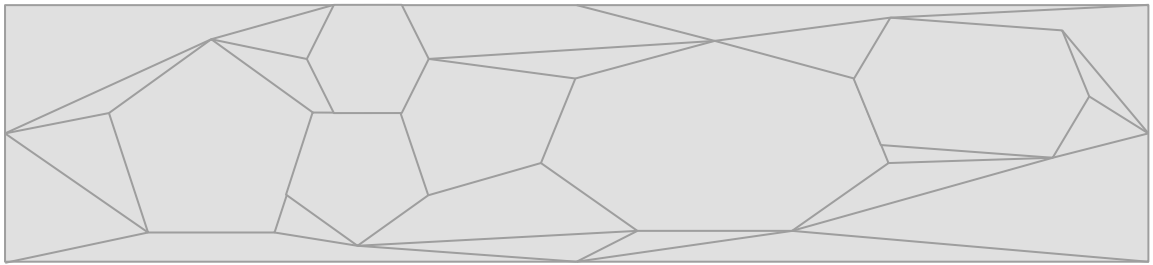
Population Avg. Loss

$$L_{pop}(\mathcal{M}, \mathcal{Z}) := \mathbb{E}_{z_i, z_l \overset{i.i.d}{\sim} \mu(\mathcal{Z})} \Big[loss[\mathcal{M}, z_i, z_l] \Big]$$

Robustness

$$\textit{if } \mathbf{x}_1, \mathbf{x}_1' \in C_i \textit{ and } \mathbf{x}_2, \mathbf{x}_2' \in C_l \textit{ such that } t_1 = t_1' = t_2 = t_2' \textit{ then}$$
$$\Big| \, loss[\mathcal{M}(\mathcal{S}_{tr}), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_{tr}), z_1, z_2] \, \Big| \leq \epsilon(\mathcal{S}_{tr}).$$

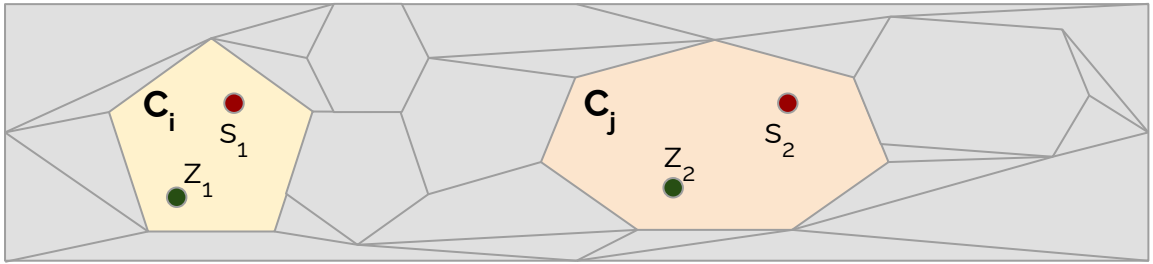
Intuitively, robustness means that for any possible units in the *support*, the loss is not far away from the loss of nearby units in *training* set, should some training units exist nearby.



Robustness

$$\textit{if } \mathbf{x}_1, \mathbf{x}_1' \in C_i \textit{ and } \mathbf{x}_2, \mathbf{x}_2' \in C_l \textit{ such that } t_1 = t_1' = t_2 = t_2' \textit{ then}$$
$$\Big| \, loss[\mathcal{M}(\mathcal{S}_{tr}), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_{tr}), z_1, z_2] \, \Big| \leq \epsilon(\mathcal{S}_{tr}).$$

Intuitively, robustness means that for any possible units in the *support*, the loss is not far away from the loss of nearby units in *training* set, should some training units exist nearby.



Multi-Robustness

$$\text{if } \mathbf{x}_1, \mathbf{x}_1' \in C_i \text{ and } \mathbf{x}_2, \mathbf{x}_2' \in C_l \text{ such that } t_1 = t_1' = t_2 = t_2' \text{ then}$$

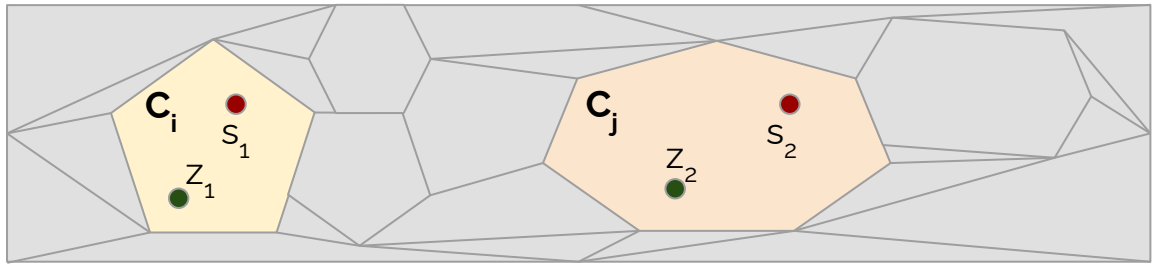
$$\left| loss[\mathcal{M}(\mathcal{S}_{tr}), s_1, s_2] - loss[\mathcal{M}(\mathcal{S}_{tr}), z_1, z_2] \right| \leq \epsilon(\mathcal{S}_{tr}).$$

$$\widehat{loss}[\mathcal{M}(\mathcal{S}_n), C_i^{(t')}, C_l^{(t')}] := \frac{1}{|C_i^{(t')}||C_l^{(t')}|} \sum_{(s_i, s_l) \in C_i^{(t')} \times C_l^{(t')}} loss[\mathcal{M}(\mathcal{S}_n), s_1, s_2]$$

$$\overline{loss}[\mathcal{M}(\mathcal{S}_n), C_i^{(t')}, C_l^{(t')}] := \mathbb{E}[loss(\mathcal{M}, Z_i, Z_l) \mid X_i' \in C_i^{(t')}, X_l' \in C_l^{(t')}]$$

Multi-robustness implies that for any two partitions of X , the *empirical* average loss over training points is not far away from the *population* average loss.

$$\forall C_i, C_l \in \boldsymbol{C}, \quad \left| \widehat{loss}[\mathcal{M}(\mathcal{S}_n), C_i^{(t')}, C_l^{(t')}] - \overline{loss}[\mathcal{M}(\mathcal{S}_n), C_i^{(t')}, C_l^{(t')}] \right| \leq \epsilon(\mathcal{S}_n)$$



Multi-Robustness

Given,

- Constant $\beta \geq 0$
- K non-empty partitions of dom(X)
  - Here, K is γ-covering number of dom(X)

Theorem 2: d_M learned by MALTS is (K,β)-multirobust with probability greater than $(1 - h(\beta,n))$

where,

- $0 \leq h(\beta,n) \leq 1$
- $h(\beta,n)$ is strictly monotonically decreasing function of β and n
- if $n \to \infty$ then $h(\beta,n) \to 0$

Multi-Robustness

Given,

- Constant $\beta \geq 0$
- K non-empty partitions of dom(X)
  - Here, K is γ-covering number of dom(X)

Theorem 2: d_M learned by MALTS is (K,β)-multirobust with probability greater than $(1 - h(\beta,n))$

where,

- $0 \leq h(\beta,n) \leq 1$
- $h(\beta,n)$ is strictly monotonically decreasing function of β and n
- if $n \to \infty$ then $h(\beta,n) \to 0$

Proof Sketch: As dom(X) is compact, the covering number is finite.
Population avg loss is equal to the expectation of empirical avg loss.
Finally, we use McDiarmids to get the high probability bound.

Generalizability

Generalizability

$$P_{\mathcal{S}_n}\left(\sum_{t' \in \mathcal{T}} \left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| \geq \epsilon \right) \leq \delta_\epsilon$$

Asymptotic Generalizability

$$\lim_{n \to \infty} \sum_{t' \in \mathcal{T}} \left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| = 0$$

With high probability, the population and empirical avg losses for the learned distance metric are close. Further, as the size of training set approaches infinity, the absolute difference of these two losses tend to zero.

Generalizability

Generalizability

$$P_{\mathcal{S}_n}\left(\sum_{t' \in \mathcal{T}} \left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| \geq \epsilon \right) \leq \delta_\epsilon$$

Asymptotic Generalizability

$$\lim_{n \to \infty} \sum_{t' \in \mathcal{T}} \left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| = 0$$

Theorem 3: d_M learned by MALTS is generalizable and asymptotically generalizable

Generalizability

Lemma 3 ***(Error Bound)*** *Given sample* $\mathcal{S}_n \overset{i.i.d}{\sim} \mu(\mathcal{Z})$ *where* $n^{(t')}$ *is the number of units with* $t_i = t'$ *in* $\mathcal{S}_n$*, and choosing* $B > 0$ *for which* $loss[\cdot, z_i, z_l] \leq B \ \forall z_i, z_l \in \mathcal{Z}$ *(*B *is finite because* $\mathcal{X}$ *is compact and* $\mathcal{Y}$ *is bounded): if a learning algorithm provides a distance metric* $\mathcal{M}(\mathcal{S}_n)$ *that is* $(K, \epsilon(\cdot))$*-multi-robust with probability* $p_{mr}(\epsilon)$*, then for any* $\mathcal{E} > 0$*, with probability greater than or equal to* $(1-\mathcal{E})(p_{mr}(\epsilon))^{K^2}$ *we have*

$$\forall t' \in \mathcal{T}, \ \left| L_{pop}(\mathcal{M}(\mathcal{S}_n), \mathcal{Z}^{(t')}) - L_{emp}(\mathcal{M}(\mathcal{S}_n), \mathcal{S}_n^{(t')}) \right| \leq \epsilon(\mathcal{S}_n^{(t')}) + 2B\sqrt{\frac{2K \ \ln(2) \ + \ 2 \ \ln(1/\mathcal{E})}{n^{(t')}}}$$

Proof Sketch:
1. Use Bretagnolle-Huber-Carol inequality bounds the empirical and population probability of units in each partition with high probability
2. Perform algebraic manipulation to use multi-robustness to bound the empirical and population avg loss for each partition.

Mixed Data Generation Process

Continuous covariates

Discrete covariates

$$\mathbf{x}_{i,p_c} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma), \quad \{x_{i,j}\}_{j \in p_d} \overset{iid}{\sim} \text{Bernoulli}(\psi), \quad \epsilon_{i,0}, \epsilon_{i,1} \overset{iid}{\sim} \mathcal{N}(0,1), \quad \epsilon_{i,\text{treat}} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$s_1, \ldots, s_{|k|} \overset{iid}{\sim} \text{Uniform}\{-1, 1\}, \quad \alpha_j | s_j \overset{iid}{\sim} \mathcal{N}(10 s_j, 9), \quad \beta_1, \ldots, \beta_{|k|} \overset{iid}{\sim} \mathcal{N}(1, 0.25)$$

Linear Baseline

Linear TE

Quadratic TE

$$\begin{aligned}
y_i^{(0)} &= \sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \epsilon_{i,0} \\
y_i^{(1)} &= \sum_{j \in k_c \cup k_d} \alpha_j x_{i,j} + \sum_{j \in k_c \cup k_d} \beta_j x_{i,j} + \sum_{j \in k_c \cup k_d} \sum_{j' \in k_c \cup k_d} x_{i,j} x_{i,j'} + \epsilon_{i,1} \\
t_i &= \mathbb{1}\left[\text{expit}\left(\sum_{j \in \kappa_c \subseteq k_c} x_{i,j} + \sum_{j \in \kappa_d \subseteq k_d} x_{i,j} - (|\kappa_c| \mu + |\kappa_d| \psi) + \epsilon_{i,\text{treat}} \right) > 0.5 \right] \\
y_i &= t_i y_i^{(1)} + (1 - t_i) y_i^{(0)}.
\end{aligned}$$