Policy learning under the f-sensitivity model

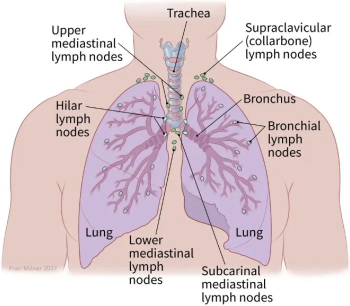
Zhimei Ren

University of Pennsylvania, Department of Statistics and Data Science

IMSI, April 2026

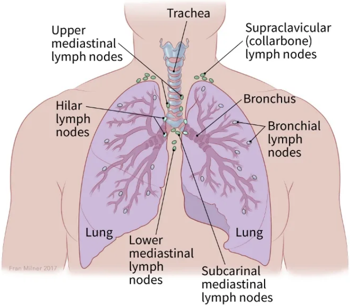
Joint work with Jingyuan Wang (NYU), Ruohan Zhan (UCL), Xinyu Liang (INSEAD), and Zhengyuan Zhou (NYU)

Non–small cell lung cancer treatment



[Source: American Cancer Society]

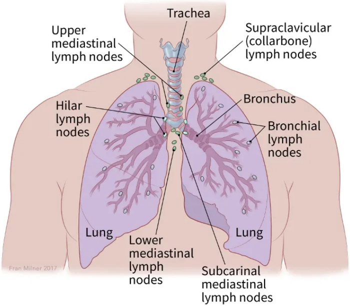
Non–small cell lung cancer treatment



[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

Non–small cell lung cancer treatment

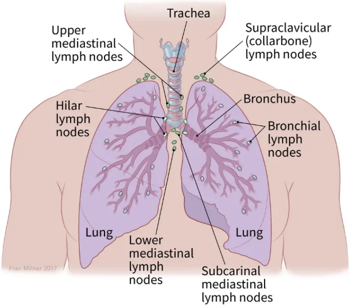


[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

- **Stage I-II:** surgical resection

Non–small cell lung cancer treatment

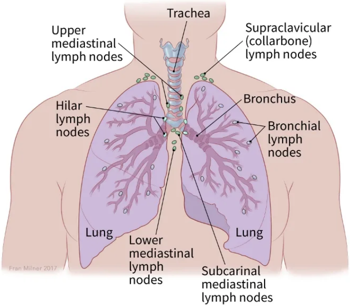


[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

- Stage I-II: surgical resection
- Stage IV: chemotherapy & non-surgical treatments

Non–small cell lung cancer treatment

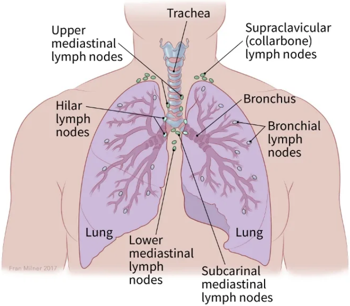


[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

- Stage I-II: surgical resection
- Stage IV: chemotherapy & non-surgical treatments
- Stage III: no clear consensus, highly heterogeneous

Non–small cell lung cancer treatment

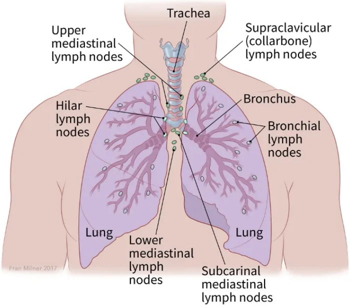


[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

- Stage I-II: surgical resection
- Stage IV: chemotherapy & non-surgical treatments
- Stage III: no clear consensus, highly heterogeneous

Non–small cell lung cancer treatment



[Source: American Cancer Society]

$\approx 85\%$ of lung cancer cases

- Stage I-II: surgical resection
- Stage IV: chemotherapy & non-surgical treatments
- Stage III: no clear consensus, highly heterogeneous

Treating stage IIIA NSCLC

The initial treatment for stage IIIA NSCLC may include some combination of radiation therapy, chemotherapy (chemo), immunotherapy, and/or surgery. For this reason, planning treatment for stage IIIA NSCLC often requires input from a medical oncologist, radiation oncologist, and a thoracic surgeon. Your treatment options depend on the size of the tumor, where it is in your lung, which lymph nodes it has spread to, your overall health, and how well you are tolerating treatment.

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

- A data set of 334,521 Stage IIIA NSCLC patients[1]

[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  $\rightsquigarrow$ quasi-random

[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  $\rightsquigarrow$ quasi-random
- Outcome: survival following treatment

[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

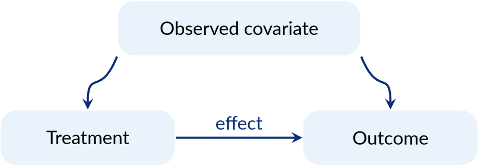
- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  $\rightsquigarrow$ quasi-random
- Outcome: survival following treatment
- Covariates: demographics, baseline health, comobidities

[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  ⇝ quasi-random
- Outcome: survival following treatment
- Covariates: demographics, baseline health, comobidities

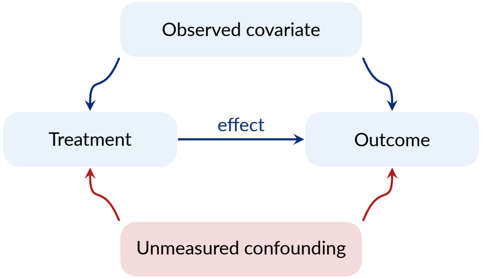


[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  $\rightsquigarrow$ quasi-random
- Outcome: survival following treatment
- Covariates: demographics, baseline health, comobidities

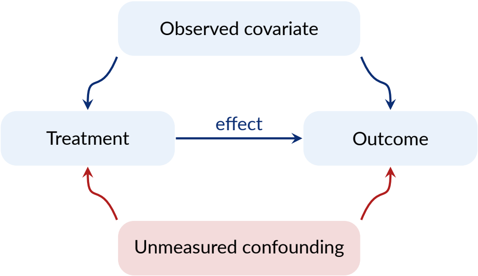


[1] curated from Truveta's real-world electronic health record (EHR) database

Treatment for stage III NSCLC

Can we learn an individualized treatment policy for stage III NSCLC from observational data?

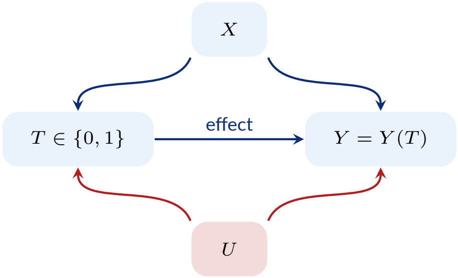
- A data set of 334,521 Stage IIIA NSCLC patients[1]
- Treatment options: surgery and chemotherapy
  $\rightsquigarrow$ quasi-random
- Outcome: survival following treatment
- Covariates: demographics, baseline health, comobidities



Common in observational studies!

[1] curated from Truveta's real-world electronic health record (EHR) database

Characterizing unmeasured confounding

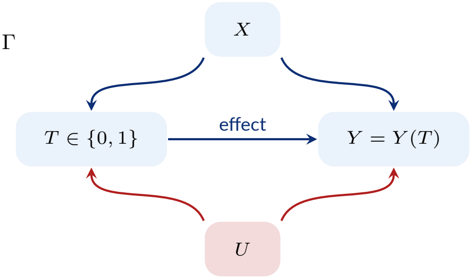


Characterizing unmeasured confounding

- Γ-sensitivity model [Rosenbaum (1987)]

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(T=1|X=x, U=u)}{\mathbb{P}(T=0|X=x, U=u)} \Big/ \frac{\mathbb{P}(T=1|X=x, U=u')}{\mathbb{P}(T=0|X=x, U=u')} \le \Gamma$$

for all x, u, u'.



$Y(0), Y(1) \rightsquigarrow$ potential outcomes

Characterizing unmeasured confounding

- Γ-sensitivity model [Rosenbaum (1987)]

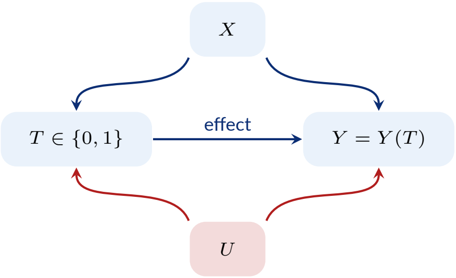
$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(T=1|X=x, U=u)}{\mathbb{P}(T=0|X=x, U=u)} \Big/ \frac{\mathbb{P}(T=1|X=x, U=u')}{\mathbb{P}(T=0|X=x, U=u')} \leq \Gamma$$

for all x, u, u'.

- marginal Γ-sensitivity model [Tan, 2006]

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(T=1|X=x, U=u)}{\mathbb{P}(T=0|X=x, U=u)} \Big/ \frac{\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=0|X=x)} \leq \Gamma$$

for all x, u.



$Y(0), Y(1) \rightsquigarrow$ potential outcomes

Characterizing unmeasured confounding

- Γ-sensitivity model [Rosenbaum (1987)]

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(T=1|X=x,U=u)}{\mathbb{P}(T=0|X=x,U=u)} \Big/ \frac{\mathbb{P}(T=1|X=x,U=u')}{\mathbb{P}(T=0|X=x,U=u')} \le \Gamma$$

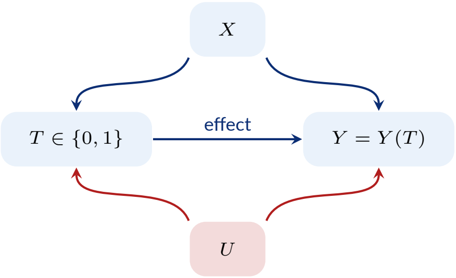
for all x, u, u'.

- marginal Γ-sensitivity model [Tan, 2006]

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(T=1|X=x,U=u)}{\mathbb{P}(T=0|X=x,U=u)} \Big/ \frac{\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=0|X=x)} \le \Gamma$$

for all x, u.

$\Gamma \ge 1$ quantifies the strength of unmeasured confounding



$Y(0), Y(1) \rightsquigarrow$ potential outcomes

How large is Γ?

- In practice, not identifiable

How large is Γ?

- In practice, not identifiable
- Assess its degree by omitting observed covariates

How large is Γ?

- In practice, not identifiable
- Assess its degree by omitting observed covariates

In NSCLC data

- Take $U =$BMI or cholesterol level

How large is Γ?

- In practice, not identifiable
- Assess its degree by omitting observed covariates

In NSCLC data

- Take U =BMI or cholesterol level
- Evidence suggests correlated with survival [Yuan et al '22, Ihara et al '24, Nie et al '24]

How large is Γ?

- In practice, not identifiable
- Assess its degree by omitting observed covariates

In NSCLC data

- Take U =BMI or cholesterol level
- Evidence suggests correlated with survival [Yuan et al '22, Ihara et al '24, Nie et al '24]

$$\mathbb{P}(T = 0 | X = x, U = u) \approx 0 \text{ for some } u \Rightarrow \Gamma = \infty!$$

How large is Γ?

- In practice, not identifiable
- Assess its degree by omitting observed covariates

In NSCLC data

- Take U =BMI or cholesterol level
- Evidence suggests correlated with survival [Yuan et al '22, Ihara et al '24, Nie et al '24]

$$\mathbb{P}(T = 0|X = x, U = u) \approx 0 \text{ for some } u \Rightarrow \Gamma = \infty!$$

Assuming the (marginal) Γ-sensitivity model will be overly conservative if the case of extreme confounding is rare

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\mathrm{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\mathrm{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

marginal sensitivity model: $\max\{\mathrm{OR}(X,U), \mathrm{OR}(X,U)^{-1}\} \le \Gamma$

f-sensitivity model: $\max\{\mathbb{E}_{U \mid X,T=1}[f(\mathrm{OR}(X,U))], \mathbb{E}_{U \mid X,T=0}[f(\mathrm{OR}(X,U)^{-1})]\} \le \rho$

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\mathsf{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

marginal sensitivity model: $\max\{\mathsf{OR}(X,U), \mathsf{OR}(X,U)^{-1}\} \leq \Gamma$

f-sensitivity model: $\max\{\mathbb{E}_{U \mid X,T=1}[f(\mathsf{OR}(X,U))], \mathbb{E}_{U \mid X,T=0}[f(\mathsf{OR}(X,U)^{-1})]\} \leq \rho$

- $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a (strongly) convex function with $f(1) = 0$

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\mathrm{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

marginal sensitivity model: $\max\{\mathrm{OR}(X,U), \mathrm{OR}(X,U)^{-1}\} \le \Gamma$

f-sensitivity model: $\max\{\mathbb{E}_{U \mid X,T=1}[f(\mathrm{OR}(X,U))], \mathbb{E}_{U \mid X,T=0}[f(\mathrm{OR}(X,U)^{-1})]\} \le \rho$

- $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a (strongly) convex function with $f(1) = 0$
- Example: $f(x) = (x-1)^2$, $f(x) = x \log x$, $f(x) = |x-1|/2$

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\mathrm{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

marginal sensitivity model: $\max\{\mathrm{OR}(X,U), \mathrm{OR}(X,U)^{-1}\} \le \Gamma$

f-sensitivity model: $\max\{\mathbb{E}_{U \mid X,T=1}[f(\mathrm{OR}(X,U))], \mathbb{E}_{U \mid X,T=0}[f(\mathrm{OR}(X,U)^{-1})]\} \le \rho$

- $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a (strongly) convex function with $f(1) = 0$
- Example: $f(x) = (x-1)^2$, $f(x) = x \log x$, $f(x) = |x-1|/2$
- $\rho \ge 0$ quantifies the strength of unmeasured confounding

The f-sensitivity model [Jin, R. and Zhou '26]

- Instead of bounding the worst-case confounding, bound the average confounding

$$\text{OR}(X,U) := \frac{\mathbb{P}(T=1|X,U)}{\mathbb{P}(T=0|X,U)} \Big/ \frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}$$

marginal sensitivity model: $\max\{\text{OR}(X,U), \text{OR}(X,U)^{-1}\} \leq \Gamma$

f-sensitivity model: $\max\{\mathbb{E}_{U \mid X, T=1}[f(\text{OR}(X,U))], \mathbb{E}_{U \mid X, T=0}[f(\text{OR}(X,U)^{-1})]\} \leq \rho$

- $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a (strongly) convex function with $f(1) = 0$
- Example: $f(x) = (x-1)^2$, $f(x) = x \log x$, $f(x) = |x-1|/2$
- $\rho \geq 0$ quantifies the strength of unmeasured confounding
- $\rho = 0 \rightsquigarrow$ no unmeasured confounding

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^{n}$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$
- Policy class of interest Π

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$
- Policy class of interest Π
- Policy value $V(\pi; Q) = \mathbb{E}_Q\left[Y(\pi(X))\right]$, where $\underbrace{(Y(1), Y(0), X) \sim Q}$

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$
- Policy class of interest Π
- Policy value $V(\pi; Q) = \mathbb{E}_Q\left[Y(\pi(X))\right]$, where $\underbrace{(Y(1), Y(0), X) \sim Q}$

w/o unmeasured confounding,

the true law of $(Y(1), Y(0), X)$ identifiable from the observed data

$$Y(1) \mid X \overset{d}{=} Y(1) \mid X, T = 1$$

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$
- Policy class of interest Π
- Policy value $V(\pi; Q) = \mathbb{E}_Q\big[Y(\pi(X))\big]$, where $\underbrace{(Y(1), Y(0), X) \sim Q}$

w/o unmeasured confounding,

the true law of $(Y(1), Y(0), X)$ identifiable from the observed data

$$Y(1) \mid X \overset{d}{=} Y(1) \mid X, T = 1$$

w/ unmeasured confounding

the true law of $(Y(1), Y(0), X)$ is not point identifiable from the observed data

$$Y(1) \mid X \overset{d}{\neq} Y(1) \mid X, T = 1$$

Policy learning

- Observe data $(X_i, T_i, Y_i)_{i=1}^n$
- Policy $\pi : \mathcal{X} \mapsto \{0, 1\}$
- Policy class of interest Π
- Policy value $V(\pi; Q) = \mathbb{E}_Q\big[Y(\pi(X))\big]$, where $\underbrace{(Y(1), Y(0), X) \sim Q}$

w/o unmeasured confounding,

the true law of $(Y(1), Y(0), X)$ identifiable from the observed data

$$Y(1) \,|\, X \overset{d}{=} Y(1) \,|\, X, T = 1$$

w/ unmeasured confounding

the true law of $(Y(1), Y(0), X)$ is is partially identifiable from the observed data under sensitivity models

$$Y(1) \,|\, X \overset{d}{\neq} Y(1) \,|\, X, T = 1$$

Partial identification under the f-sensitivity model

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, there is

$$D_f\big(\underbrace{P_{Y(t)|X=x,T=t}}_{\text{observable}} \,\|\, P_{Y(t)|X=x}\big) \le \rho, \quad t = 0, 1$$

Partial identification under the f-sensitivity model

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, there is

$$D_f\big(\underbrace{P_{Y(t)|X=x,T=t}}_{\text{observable}} \,\|\, P_{Y(t)|X=x}\big) \le \rho, \quad t = 0, 1$$

- $D_f(P\|Q)$ denotes the f-divergence between distributions P and Q

Partial identification under the f-sensitivity model

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, there is

$$D_f\big(\underbrace{P_{Y(t)\mid X=x,T=t}}_{\text{observable}} \,\|\, P_{Y(t)\mid X=x}\big) \le \rho, \quad t = 0,1$$

- $D_f(P\|Q)$ denotes the f-divergence between distributions P and Q
- Example: $f = (x-1)^2 \rightsquigarrow \chi^2$-divergence; $f(x) = x \log x \rightsquigarrow$ KL-divergence

Partial identification under the f-sensitivity model

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, there is

$$D_f\big(\underbrace{P_{Y(t)\mid X=x,T=t}}_{\text{observable}} \,\|\, P_{Y(t)\mid X=x}\big) \le \rho, \quad t = 0, 1$$

- $D_f(P\|Q)$ denotes the f-divergence between distributions P and Q
- Example: $f = (x-1)^2 \rightsquigarrow \chi^2$-divergence; $f(x) = x \log x \rightsquigarrow$ KL-divergence
- Learning objective: find the optimal policy maximizing the worst-case value

$$\pi^* = \operatorname*{argmin}_{\pi \in \Pi} \mathcal{V}_\rho(\pi) = \operatorname*{argmax}_{\pi \in \Pi} \min_{Q \in \mathcal{Q}_f(\rho)} V(\pi; Q),$$

where $\mathcal{Q}_f(\rho)$ is the set of laws of $(Y(1), Y(0), X)$ with

Partial identification under the f-sensitivity model

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, there is

$$D_f\big(\underbrace{P_{Y(t)|X=x,T=t}}_{\text{observable}} \,\|\, P_{Y(t)|X=x}\big) \le \rho, \quad t = 0, 1$$

- $D_f(P\|Q)$ denotes the f-divergence between distributions P and Q
- Example: $f = (x-1)^2 \rightsquigarrow \chi^2$-divergence; $f(x) = x \log x \rightsquigarrow$ KL-divergence
- Learning objective: find the optimal policy maximizing the worst-case value

$$\pi^* = \operatorname*{argmin}_{\pi \in \Pi} \mathcal{V}_\rho(\pi) = \operatorname*{argmax}_{\pi \in \Pi} \min_{Q \in \mathcal{Q}_f(\rho)} V(\pi; Q),$$

where $\mathcal{Q}_f(\rho)$ is the set of laws of $(Y(1), Y(0), X)$ with

$$P_X = P_X^{\text{obs}}, \quad D_f(P^{\text{obs}}_{Y(t)|X=x,T=t} \| P_{Y(t)|X=x}) \le \rho, \ \forall t = 0, 1$$

Offline data collection mechanism

How is $\mathcal{D} = \{X_i, T_i, Y_i\}_{i=1}^n$ collected? Consider two types of data collection mechanisms:

I.i.d. data

- $(U_i, X_i, T_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} P, i = 1, \dots, n$
- Observe (X_i, T_i, Y_i) with $Y_i = Y_i(T_i)$
- Propensity score $e(X) = \mathbb{P}(T = 1|X) \in [\epsilon, 1 - \epsilon]$

Adaptively collected data

- $(U_i, X_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} P, i = 1, \dots, n$
- Intended treatment $T_i \sim$ Bern$(e(X_i, \mathcal{H}_{i-1}))$
- Actual treatment T_i^* may $\neq T_i$ (non-compliance)
- Observe $(X_i, T_i, , T_i^*, Y_i)$ with $Y_i = Y_i(T_i^*)$
- $T_i^* \mid X_i, T_i$ is time-invariant

Offline data collection mechanism

How is $\mathcal{D} = \{X_i, T_i, Y_i\}_{i=1}^n$ collected? Consider two types of data collection mechanisms:

I.i.d. data

- $(U_i, X_i, T_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} P, i = 1, \ldots, n$
- Observe (X_i, T_i, Y_i) with $Y_i = Y_i(T_i)$
- Propensity score $e(X) = \mathbb{P}(T = 1|X) \in [\epsilon, 1 - \epsilon]$

Adaptively collected data

- $(U_i, X_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} P, i = 1, \ldots, n$
- Intended treatment $T_i \sim$ Bern$(e(X_i, \mathcal{H}_{i-1}))$
- Actual treatment T_i^* may $\neq T_i$ (non-compliance)
- Observe $(X_i, T_i, , T_i^*, Y_i)$ with $Y_i = Y_i(T_i^*)$
- $T_i^* \mid X_i, T_i$ is time-invariant

Distributionally robust policy learning

Learning objective:

$$\pi^* = \operatorname*{argmax}_{\pi \in \Pi} \mathcal{V}_\rho(\pi),$$

Distributionally robust policy learning

Learning objective:

$$\pi^* = \operatorname*{argmax}_{\pi \in \Pi} \mathcal{V}_\rho(\pi),$$

Evaluation metric: for any policy $\hat{\pi}$ obtained from $\mathcal{D}$

$$\text{regret}(\hat{\pi}) = \mathcal{V}_\rho(\pi^*) - \mathcal{V}_\rho(\hat{\pi})$$

Distributionally robust policy learning

Learning objective:

$$\pi^* = \underset{\pi \in \Pi}{\text{argmax}}\ \mathcal{V}_\rho(\pi),$$

Evaluation metric: for any policy $\hat{\pi}$ obtained from $\mathcal{D}$

$$\text{regret}(\hat{\pi}) = \mathcal{V}_\rho(\pi^*) - \mathcal{V}_\rho(\hat{\pi})$$

Construct $\hat{\mathcal{V}}_\rho(\pi)$ → Optimize $\hat{\mathcal{V}}_\rho(\pi)$ over $\pi \in \Pi$ → Output robust policy $\hat{\pi}$

Policy evaluation: from primal to dual

Fix $\pi \in \Pi$

- Need to estimate the policy value under the worst-case law of $Y(1), Y(0) \mid X$

$$\begin{aligned}
\mathcal{V}_\rho(\pi) &= \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(\pi(X))] \\
&= \mathbb{E}_{P_X}\left[\pi(X) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(1)|X] + (1-\pi(X)) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(0)|X]\right]
\end{aligned}$$

Policy evaluation: from primal to dual

Fix $\pi \in \Pi$

- Need to estimate the policy value under the worst-case law of $Y(1), Y(0) \mid X$

$$
\begin{aligned}
\mathcal{V}_\rho(\pi) &= \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(\pi(X))] \\
&= \mathbb{E}_{P_X}\left[\pi(X) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(1)|X] + (1-\pi(X)) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(0)|X]\right]
\end{aligned}
$$

- Easier with the dual problem!

Policy evaluation: from primal to dual

Fix $\pi \in \Pi$

- Need to estimate the policy value under the worst-case law of $Y(1), Y(0) \mid X$

$$\begin{aligned}\mathcal{V}_\rho(\pi) &= \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(\pi(X))] \\ &= \mathbb{E}_{P_X}\left[\pi(X) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(1)|X] + (1-\pi(X)) \cdot \min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(0)|X]\right]\end{aligned}$$

- Easier with the dual problem!

Lemma. [Jin, R., Zhou '26]

Under the f-sensitivity model with parameter ρ, for any $t = 0, 1$,

$$\min_{Q \in \mathcal{Q}_f(\rho)} \mathbb{E}_Q[Y(t)|X] = -\min_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E}_P\left[\alpha f^*\left(-\frac{Y(t)+\eta}{\alpha}\right) + \eta + \alpha\rho \,\middle|\, X, T = t\right],$$

where $f^*(z) = \sup_{x \geq 0}\{zx - f(x)\}$ is the convex conjugate of f.

Policy evaluation: estimator

Dual opt. problem: $$\min_{\alpha\geq 0,\eta\in\mathbb{R}} \mathbb{E}_P\left[\underbrace{\alpha f^*\Big(-\frac{Y(t)+\eta}{\alpha}\Big)+\eta+\alpha\rho}_{R_t(\alpha,\rho)}\,\Big|\, X, T=t\right]$$

Policy evaluation: estimator

Dual opt. problem: $$\min_{\alpha\geq 0,\eta\in\mathbb{R}} \mathbb{E}_P\left[\underbrace{\alpha f^*\Big(-\frac{Y(t)+\eta}{\alpha}\Big)+\eta+\alpha\rho}_{R_t(\alpha,\rho)}\,\Big|\, X, T=t\right]$$

- Plug-in approach?

$$\hat{\mathcal{V}}_\rho(\pi)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}\{T_i=\pi(X_i)\}}{\hat{e}(X_i)}\cdot R_{\pi(X_i)}(\hat{\alpha}(X_i),\hat{\eta}(X_i))\right]$$

Policy evaluation: estimator

Dual opt. problem: $$\min_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E}_P \left[\underbrace{\alpha f^* \left(- \frac{Y(t) + \eta}{\alpha} \right) + \eta + \alpha \rho}_{R_t(\alpha, \rho)} \,\middle|\, X, T = t \right]$$

- Plug-in approach?

$$\hat{\mathcal{V}}_\rho(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{1}\{T_i = \pi(X_i)\}}{\hat{e}(X_i)} \cdot R_{\pi(X_i)}(\hat{\alpha}(X_i), \hat{\eta}(X_i)) \right]$$

Policy evaluation: estimator

Dual opt. problem: $\min\limits_{\alpha \geq 0, \eta \in \mathbb{R}} \mathbb{E}_P \Big[\underbrace{\alpha f^* \Big(-\frac{Y(t) + \eta}{\alpha} \Big) + \eta + \alpha \rho}_{R_t(\alpha, \rho)} \Big| X, T = t \Big]$

- Plug-in approach?

$$\hat{\mathcal{V}}_\rho(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{1}\{T_i = \pi(X_i)\}}{\hat{e}(X_i)} \cdot R_{\pi(X_i)}\big(\hat{\alpha}(X_i), \hat{\eta}(X_i) \big) \right]$$

- Nuisance functions $\hat{e}(\cdot), \hat{\alpha}(\cdot), \hat{\eta}(\cdot)$ are estimated by the same data $\mathcal{D} \Rightarrow$ inefficient!

Policy evaluation: cross-fitting estimator

Fold 1 — Solve dual opt. obtain $\hat{\alpha}_t^{(1)}(x), \hat{\eta}_t^{(1)}(x)$

Fold 2 — Estimate $e(x)$ and $r_t := \mathbb{E}\big[R_t(\hat{\alpha}^{(1)}, \hat{\eta}^{(1)}) \mid X\big] \rightsquigarrow (\hat{e}^{(1)}, \hat{r}_t^{(1)})$

Fold 3 — $\hat{\mathcal{V}}_\rho^{(1)}(\pi) = \sum_{i \in \mathcal{D}_3} \frac{\mathbb{1}\{T_i = \pi(X_i)\}}{\hat{e}^{(1)}(\pi(X_i))} \Big(R_{\pi(X_i)}(\hat{\alpha}^{(1)}, \hat{\eta}^{(1)}) - \hat{r}^{(1)}(X_i) \Big) + \hat{r}^{(1)}(X_i)$

Policy evaluation: cross-fitting estimator

Fold 1 $\hat{\mathcal{V}}_\rho^{(2)}(\pi) = \sum_{i \in \mathcal{D}_1} \frac{\mathbb{1}\{T_i = \pi(X_i)\}}{\hat{e}^{(2)}(\pi(X_i))} \Big(R_{\pi(X_i)}(\hat{\alpha}^{(2)}, \hat{\eta}^{(2)}) - \hat{r}^{(2)}(X_i) \Big) + \hat{r}^{(2)}(X_i)$

Fold 2 Solve dual opt. obtain $\hat{\alpha}_t^{(2)}(x), \hat{\eta}_t^{(2)}(x)$

Fold 3 Estimate $e(x)$ and $r_t := \mathbb{E}\big[R_t(\hat{\alpha}^{(2)}, \hat{\eta}^{(2)}) \mid X\big] \rightsquigarrow (\hat{e}^{(2)}, \hat{r}_t^{(2)})$

Policy evaluation: cross-fitting estimator

Fold 1 Estimate $e(x)$ and $r_t := \mathbb{E}\big[R_t(\hat{\alpha}^{(3)}, \hat{\eta}^{(3)}) \mid X\big] \rightsquigarrow (\hat{e}^{(3)}, \hat{r}_t^{(3)})$

Fold 2 $\hat{\mathcal{V}}_\rho^{(3)}(\pi) = \sum_{i \in \mathcal{D}_2} \frac{\mathbb{1}\{T_i = \pi(X_i)\}}{\hat{e}^{(3)}(\pi(X_i))} \Big(R_{\pi(X_i)}(\hat{\alpha}^{(3)}, \hat{\eta}^{(3)}) - \hat{r}^{(3)}(X_i)\Big) + \hat{r}^{(3)}(X_i)$

Fold 3 Solve dual opt. obtain $\hat{\alpha}_t^{(3)}(x), \hat{\eta}_t^{(3)}(x)$

Policy evaluation: cross-fitting estimator

$$\hat{\mathcal{V}}_\rho(\pi) = -\frac{1}{3}\sum_{k=1}^{3}\hat{\mathcal{V}}^{(k)}$$

Policy evaluation: cross-fitting estimator

$$\hat{\mathcal{V}}_\rho(\pi) = -\frac{1}{3}\sum_{k=1}^{3}\hat{\mathcal{V}}^{(k)}$$

Theorem (Double robustness of the policy value estimator). [Wang, R., Zhan, Liang, Zhou '26+]

Under mild regularity conditions, suppose for each fold $k \in [3]$

$$\|\hat{e}^{(k)} - e\|_2 = o_p(n^{-\gamma_1}), \quad \|\hat{r}_t^{(k)} - r_t\|_2 = o_p(n^{-\gamma_2}), \quad t = 0, 1$$
$$\|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_2 = o_p(n^{-1/4}), \quad \|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_\infty = o_p(1),$$

for some $\gamma_1, \gamma_2 > 0$ satisfying $\gamma_1 + \gamma_2 > 1/2$. Then for any π,

$$\sqrt{n}(\hat{\mathcal{V}}_\rho(\pi) - \mathcal{V}_\rho(\pi)) \Rightarrow \mathcal{N}(0, \sigma_\pi^2).$$

Policy optimization

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \hat{\mathcal{V}}_\rho(\pi)$$

Policy optimization

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \hat{\mathcal{V}}_{\rho}(\pi)$$

- Inner opt. problem only need to solved for $t = 0, 1$

Policy optimization

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \hat{\mathcal{V}}_\rho(\pi)$$

- Inner opt. problem only need to solved for $t = 0, 1$
- Efficient implementation for tree-based policy class via dynamic programming

Policy optimization: regret upper bound

Theorem (Regret upper bound). [Wang, R., Zhan, Liang, Zhou '26+]

Under mild regularity conditions, suppose for each fold $k \in [3]$

$$\|\hat{e}^{(k)} - e\|_2 = o_p(n^{-\gamma_1}), \quad \|\hat{r}_t^{(k)} - r_t\|_2 = o_p(n^{-\gamma_2}), \quad t = 0, 1$$
$$\|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_2 = o_p(n^{-1/4}), \quad \|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_\infty = o_p(1),$$

for some $\gamma_1, \gamma_2 > 0$ satisfying $\gamma_1 + \gamma_2 > 1/2$. Then for any $\beta \in (0, 1)$ and n sufficiently large,

$$\text{regret}(\hat{\pi}) \leq O\left(\frac{\kappa(\Pi) + \log(1/\beta)}{\sqrt{n}}\right)$$

Policy optimization: regret upper bound

Theorem (Regret upper bound). [Wang, R., Zhan, Liang, Zhou '26+]

Under mild regularity conditions, suppose for each fold $k \in [3]$

$$\|\hat{e}^{(k)} - e\|_2 = o_p(n^{-\gamma_1}), \quad \|\hat{r}_t^{(k)} - r_t\|_2 = o_p(n^{-\gamma_2}), \quad t = 0, 1$$
$$\|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_2 = o_p(n^{-1/4}), \quad \|(\hat{\alpha}^{(k)}, \hat{\eta}^{(k)}) - (\alpha, \eta)\|_\infty = o_p(1),$$

for some $\gamma_1, \gamma_2 > 0$ satisfying $\gamma_1 + \gamma_2 > 1/2$. Then for any $\beta \in (0, 1)$ and n sufficiently large,

$$\text{regret}(\hat{\pi}) \le O\left(\frac{\kappa(\Pi) + \log(1/\beta)}{\sqrt{n}}\right)$$

- $\kappa(\Pi)$: entropy integral of Π under the Hamming distance $\rightsquigarrow$ complexity of Π

Policy optimization: regret lower bound

Theorem (Minimax lower bound). [Wang, R., Zhan, Liang, Zhou '26+]

Suppose $\rho \leq 0.2$ and $n \geq \text{VCdim}(\Pi)^2$. There exists a distribution over $(X, U, T, Y(0), Y(1))$ satisfying the same condition in the upper bound result, such that for any $\hat{\pi}$,

$$\mathbb{E}\left[\text{regret}(\hat{\pi})\right] \geq \Omega\left(\sqrt{\frac{\text{VCdim}(\Pi)}{n}}\right)$$

Policy optimization: regret lower bound

Theorem (Minimax lower bound). [Wang, R., Zhan, Liang, Zhou '26+]

Suppose $\rho \leq 0.2$ and $n \geq \mathrm{VCdim}(\Pi)^2$. There exists a distribution over $(X, U, T, Y(0), Y(1))$ satisfying the same condition in the upper bound result, such that for any $\hat{\pi}$,

$$\mathbb{E}\left[\mathrm{regret}(\hat{\pi})\right] \geq \Omega\left(\sqrt{\frac{\mathrm{VCdim}(\Pi)}{n}}\right)$$

- $\mathrm{VCdim}(\Pi)$: the VC dimension of Π

Policy optimization: regret lower bound

Theorem (Minimax lower bound). [Wang, R., Zhan, Liang, Zhou '26+]

Suppose $\rho \leq 0.2$ and $n \geq \mathsf{VCdim}(\Pi)^2$. There exists a distribution over $(X, U, T, Y(0), Y(1))$ satisfying the same condition in the upper bound result, such that for any $\hat{\pi}$,

$$\mathbb{E}\left[\mathsf{regret}(\hat{\pi})\right] \geq \Omega\left(\sqrt{\frac{\mathsf{VCdim}(\Pi)}{n}}\right)$$

- $\mathsf{VCdim}(\Pi)$: the VC dimension of Π
- $\sqrt{\mathsf{VCdim}} = \Omega(\kappa(\Pi)) \Rightarrow$ matching lower bound!

Policy optimization: regret lower bound

Theorem (Minimax lower bound). [Wang, R., Zhan, Liang, Zhou '26+]

Suppose $\rho \le 0.2$ and $n \ge \mathrm{VCdim}(\Pi)^2$. There exists a distribution over $(X, U, T, Y(0), Y(1))$ satisfying the same condition in the upper bound result, such that for any $\hat{\pi}$,

$$\mathbb{E}\left[\mathrm{regret}(\hat{\pi})\right] \ge \Omega\left(\sqrt{\frac{\mathrm{VCdim}(\Pi)}{n}}\right)$$

- $\mathrm{VCdim}(\Pi)$: the VC dimension of Π
- $\sqrt{\mathrm{VCdim}} = \Omega(\kappa(\Pi)) \Rightarrow$ matching lower bound!

When $\mathrm{Err}(e) \times \mathrm{Err}(r_t) = O(n^{-1/2})$ and $\mathrm{Err}(\alpha) \vee \mathrm{Err}(\eta) = O(n^{-1/4})$,

Minimax optimal!

When data are collected adaptively

Main distinction

When data are collected adaptively

Main distinction

- Propensity estimation:

$$\mathbb{P}(T_i^* = t \mid X_i, \mathcal{H}_{i-1}) = \underbrace{e(X_i, \mathcal{H}_{i-1})}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 1, X_i) + \underbrace{(1 - e(X_i, \mathcal{H}_{i-1}))}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 0, X_i)$$

When data are collected adaptively

Main distinction

- Propensity estimation:

$$\mathbb{P}(T_i^* = t \mid X_i, \mathcal{H}_{i-1}) = \underbrace{e(X_i, \mathcal{H}_{i-1})}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 1, X_i) + \underbrace{(1 - e(X_i, \mathcal{H}_{i-1}))}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 0, X_i)$$

When data are collected adaptively

Main distinction

- Propensity estimation:

$$\mathbb{P}(T_i^* = t \mid X_i, \mathcal{H}_{i-1}) = \underbrace{e(X_i, \mathcal{H}_{i-1})}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 1, X_i) + \underbrace{(1 - e(X_i, \mathcal{H}_{i-1}))}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 0, X_i)$$

When data are collected adaptively

Main distinction

- Propensity estimation:

$$\mathbb{P}(T_i^* = t \mid X_i, \mathcal{H}_{i-1}) = \underbrace{e(X_i, \mathcal{H}_{i-1})}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 1, X_i) + \underbrace{(1 - e(X_i, \mathcal{H}_{i-1}))}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 0, X_i)$$

- Accommodates unbalanced sampling: $e(X_i; \mathcal{H}_{i-1}) \geq \epsilon_i$

When data are collected adaptively

Main distinction

- Propensity estimation:

$$\mathbb{P}(T_i^* = t \mid X_i, \mathcal{H}_{i-1}) = \underbrace{e(X_i, \mathcal{H}_{i-1})}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 1, X_i) + \underbrace{(1 - e(X_i, \mathcal{H}_{i-1}))}_{\text{known}} \cdot \mathbb{P}(T_i^* = t \mid T_i = 0, X_i)$$

- Accommodates unbalanced sampling: $e(X_i; \mathcal{H}_{i-1}) \geq \epsilon_i$

Similar algorithm (w/o cross-fitting)

$$\mathbb{E}[\text{regret}(\hat{\pi})] = O\Big(\frac{\kappa(\Pi)}{n}\sqrt{\sum_i \frac{1}{\epsilon_i^2}}\Big)$$

NSCLC dataset revisited

NSCLC dataset revisited

- U =BMI or U =cholesterol level

NSCLC dataset revisited

- U =BMI or U =cholesterol level
- Half of data $\mathcal{D}_1$ w/ full set of covariates: estimate the counterfactual (for evaluation)

NSCLC dataset revisited

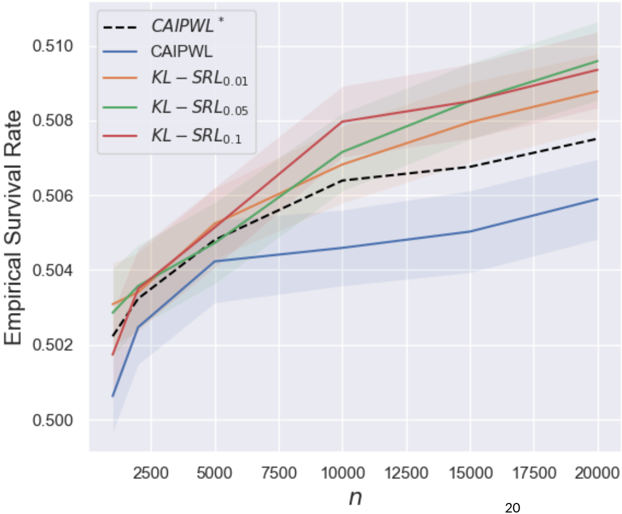
Patient covariates $\rightarrow$ Surgery or Chemotherapy? $\rightarrow$ Survival?

- U =BMI or U =cholesterol level
- Half of data $\mathcal{D}_1$ w/ full set of covariates: estimate the counterfactual (for evaluation)
- The other half $\mathcal{D}_2$ w/o BMI or cholesterol level: learn a robust policy $\hat{\pi}$ and evaluate it

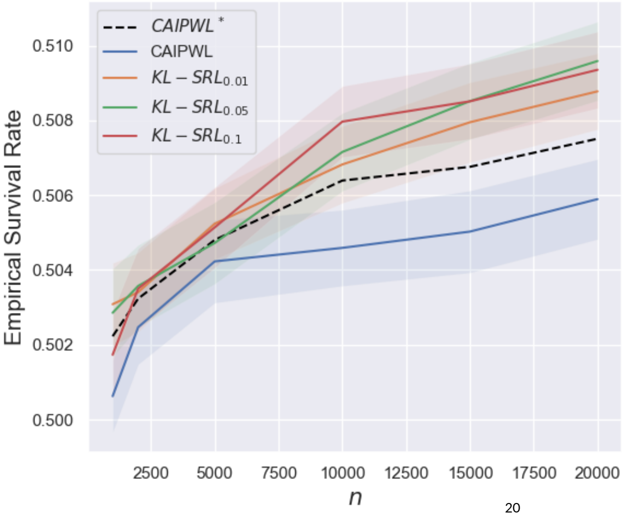
NSCLC dataset revisited

- U =BMI or U =cholesterol level
- Half of data $\mathcal{D}_1$ w/ full set of covariates: estimate the counterfactual (for evaluation)
- The other half $\mathcal{D}_2$ w/o BMI or cholesterol level: learn a robust policy $\hat{\pi}$ and evaluate it
- Choice of ρ? Hard to estimate in general $\Rightarrow$ conduct sensitivity analysis by varying ρ

BMI as unmeasured confounder

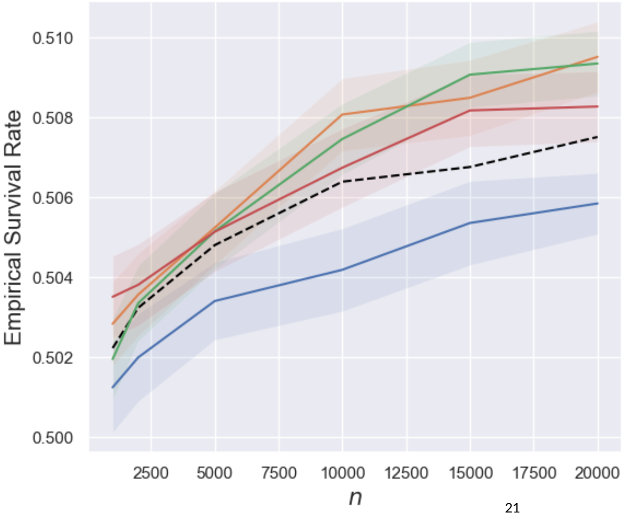


BMI as unmeasured confounder

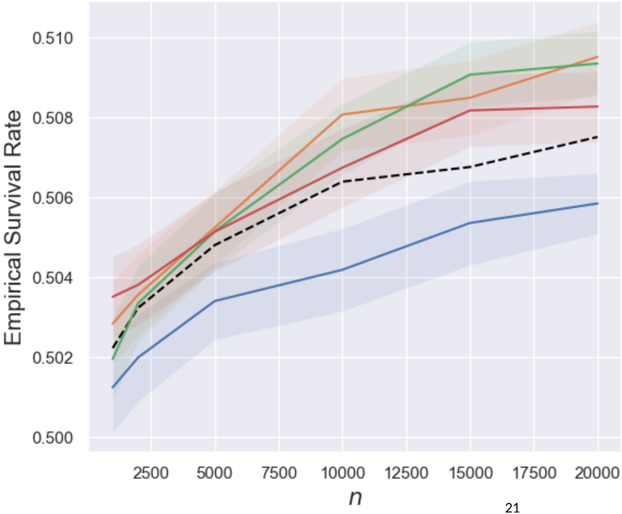


- $f(x) = x \log x \rightsquigarrow$ KL divergence
- Baseline: CAIPWL
  [Zhou, Athey and Wager '22]
- Oracle baseline: CAIPWL
  with BMI observed

Cholesterol level as unmeasured confounder

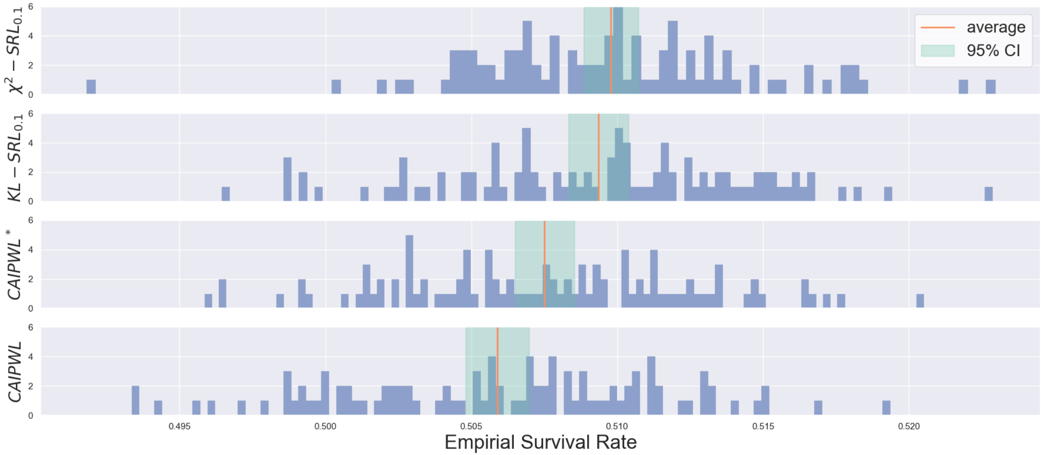


Cholesterol level as unmeasured confounder



- $f(x) = x \log x \rightsquigarrow$ KL divergence
- Baseline: CAIPWL
  [Zhou, Athey and Wager '22]
- Oracle baseline: CAIPWL
  with cholesterol level observed

Histogram of survival rate



Summary

- The f-sensitivity model: a dsensitivity model measuring the average strength of unmeasured confounding

Summary

- The f-sensitivity model: a dsensitivity model measuring the average strength of unmeasured confounding
- Efficient policy evaluation and optimization algorithms under the f-sensitivity model

Summary

- The f-sensitivity model: a dsensitivity model measuring the average strength of unmeasured confounding
- Efficient policy evaluation and optimization algorithms under the f-sensitivity model
- Applies to both i.i.d. and adaptively collected data

Summary

- The f-sensitivity model: a dsensitivity model measuring the average strength of unmeasured confounding
- Efficient policy evaluation and optimization algorithms under the f-sensitivity model
- Applies to both i.i.d. and adaptively collected data
- Empirical results on NSCLC dataset suggest the robust policy can outperform the baseline

Related work:

- Ying Jin, Zhimei Ren, and Zhengyuan Zhou. "Sensitivity Analysis Under the f-Sensitivity Model: A Distributional Robustness Perspective." *Operations Research* (2026).
- Jinyuan Wang, Zhimei Ren, Ruohan Zhan, and Zhengyuan Zhou. "Distributionally robust policy learning under concept drifts." *ICML* (2025).
- Jinyuan Wang, Zhimei Ren, Ruohan Zhan, Xinyu Liang and Zhengyuan Zhou. "Optimal policy learning under the f-sensitivity model." *in preparation* (2026+).

Thank you!