Fair comparisons of causal parameters with many treatments and positivity violations

Alec McClean, Yiting Li, Sunjae Bae,
Mara A. McAdams-DeMarco, Iván Díaz, and Wenbo Wu

New York University Grossman School of Medicine

https://arxiv.org/abs/2410.13522

Motivation and background

Provider profiling ≡ observational data with multi-valued treatment

Provider profiling: compare healthcare providers in terms of patient outcomes

$Z_i = \{X_i, A_i, Y_i\}$ for $i = 1, \ldots, n$ with:

- $X \in \mathbb{R}^p$: covariates (e.g., patient information)
- $A \in \{1, \ldots, d\}$: multi-valued treatment (e.g., provider)
- $Y \in \mathbb{R}$: outcome (e.g., mortality, 30-day unplanned readmission)

Goal: "fair" comparisons of treatment efficacy

Method:

1. Estimate a (causal) parameter for each treatment a: ψ_a
2. Create league table from $\{\psi_1, \ldots, \psi_d\}$.

Provider profiling ≡ observational data with multi-valued treatment

Provider profiling: compare healthcare providers in terms of patient outcomes

$Z_i = \{X_i, A_i, Y_i\}$ for $i = 1, \ldots, n$ with:

- $X \in \mathbb{R}^p$: covariates (e.g., patient information)
- $A \in \{1, \ldots, d\}$: multi-valued treatment (e.g., provider)
- $Y \in \mathbb{R}$: outcome (e.g., mortality, 30-day unplanned readmission)

Goal: "fair" comparisons of treatment efficacy
Method:

1. Estimate a (causal) parameter for each treatment a: ψ_a
2. Create league table from $\{\psi_1, \ldots, \psi_d\}$.

Provider profiling ≡ observational data with multi-valued treatment

Provider profiling: compare healthcare providers in terms of patient outcomes

$Z_i = \{X_i, A_i, Y_i\}$ for $i = 1, \ldots, n$ with:

- $X \in \mathbb{R}^p$: covariates (e.g., patient information)
- $A \in \{1, \ldots, d\}$: multi-valued treatment (e.g., provider)
- $Y \in \mathbb{R}$: outcome (e.g., mortality, 30-day unplanned readmission)

Goal: "fair" comparisons of treatment efficacy

Method:

1. Estimate a (causal) parameter for each treatment a: ψ_a
2. Create league table from $\{\psi_1, \ldots, \psi_d\}$.

Provider profiling $\equiv$ observational data with multi-valued treatment

Provider profiling: compare healthcare providers in terms of patient outcomes

$Z_i = \{X_i, A_i, Y_i\}$ for $i = 1, \ldots, n$ with:

- $X \in \mathbb{R}^p$: covariates (e.g., patient information)
- $A \in \{1, \ldots, d\}$: multi-valued treatment (e.g., provider)
- $Y \in \mathbb{R}$: outcome (e.g., mortality, 30-day unplanned readmission)

Goal: "fair" comparisons of treatment efficacy
Method:

1. Estimate a (causal) parameter for each treatment a: ψ_a
2. Create league table from $\{\psi_1, \ldots, \psi_d\}$.

Treatment-specific mean (TSM):

$\psi_a = \mathbb{E}(Y^a)$ (average potential outcome if all took treatment a).

TSMs allow "fair" comparisons since the distribution of covariates is held constant across treatments.

Same set of patients *would* attend provider a in $\mathbb{E}(Y^a)$ as *would* attend provider b in $\mathbb{E}(Y^b)$ (all patients)

TSMs are fair

Treatment-specific mean (TSM):

$\psi_a = \mathbb{E}(Y^a)$ (average potential outcome if all took treatment a).

TSMs allow "fair" comparisons since the distribution of covariates is held constant across treatments.

Same set of patients *would* attend provider a in $\mathbb{E}(Y^a)$ as *would* attend provider b in $\mathbb{E}(Y^b)$ (all patients)

Treatment-specific mean (TSM):

$\psi_a = \mathbb{E}(Y^a)$ (average potential outcome if all took treatment a).

TSMs allow "fair" comparisons since the distribution of covariates is held constant across treatments.

Same set of patients *would* attend provider a in $\mathbb{E}(Y^a)$ as *would* attend provider b in $\mathbb{E}(Y^b)$ (all patients)

TSMs are fair

Treatment-specific mean (TSM):

$\psi_a = \mathbb{E}(Y^a)$ (average potential outcome if all took treatment a).

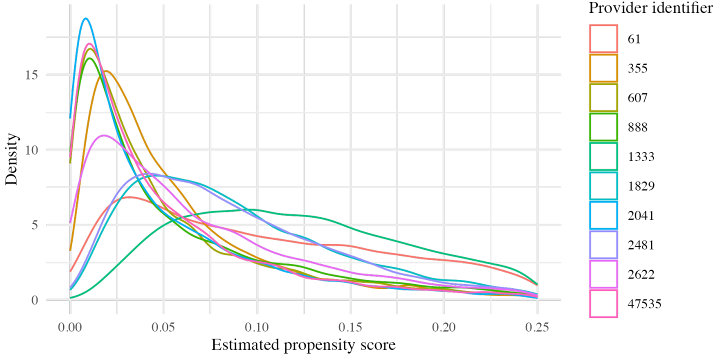
TSMs allow "fair" comparisons since the distribution of covariates is held constant across treatments.

Same set of patients *would* attend provider a in $\mathbb{E}(Y^a)$ as *would* attend provider b in $\mathbb{E}(Y^b)$ (all patients)

TSMs are unreasonable target because of positivity violations

But: With many treatments, positivity violations
$\rightarrow \pi_a(X) = \mathbb{P}(A = a \mid X) = 0$ or ≈ 0

$\mathbb{E}(Y^a)$ unidentifiable or estimators with have high variance.

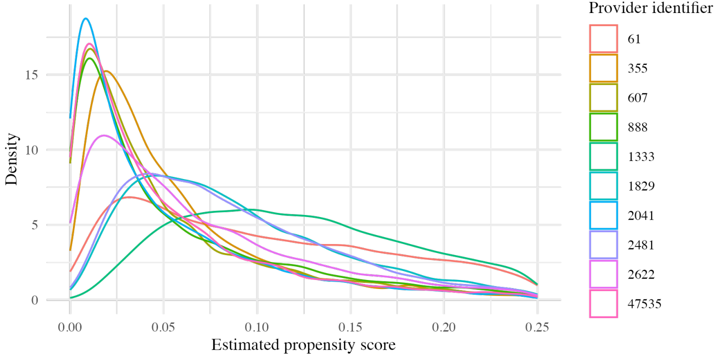


10 largest providers in NY state and their patients

TSMs are unreasonable target because of positivity violations

But: With many treatments, positivity violations
$\rightarrow \pi_a(X) = \mathbb{P}(A = a \mid X) = 0$ or ≈ 0

$\mathbb{E}(Y^a)$ unidentifiable or estimators with have high variance.

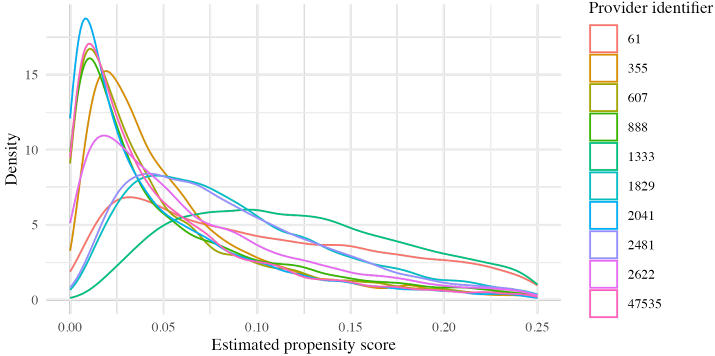


10 largest providers in NY state and their patients

TSMs are unreasonable target because of positivity violations

But: With many treatments, positivity violations
$\rightarrow \pi_a(X) = \mathbb{P}(A = a \mid X) = 0$ or ≈ 0

$\mathbb{E}(Y^a)$ unidentifiable or estimators with have high variance.



10 largest providers in NY state and their patients

Alternative: dynamic stochastic ("soft") interventions, but care is necessary

Dynamic stochastic intervention targeting treatment a:
For $b \in \{1, \ldots, d\}$, define an *interventional propensity score*

$$q_a(A = b \mid V) \quad \text{(a function of } V \subseteq X\text{)}.$$

The parameter targeting treatment a is

$$\psi_a = \mathbb{E}\left(Y^{Q_a}\right),$$

where $Q_a \sim \text{Categorical}\left\{q_a(A = 1 \mid V), \ldots, q_a(A = d \mid V)\right\}$.
These can adapt to positivity violations!

Issue: To respect positivity, may have non-zero probability of attending many treatments, and so $\psi_a - \psi_b$ can be driven by performance at treatment c. **Unfair?**

More generally: what does fairness mean?

Alternative: dynamic stochastic ("soft") interventions, but care is necessary

Dynamic stochastic intervention targeting treatment a:
For $b \in \{1, \ldots, d\}$, define an *interventional propensity score*

$$q_a(A = b \mid V) \quad \text{(a function of } V \subseteq X\text{).}$$

The parameter targeting treatment a is

$$\psi_a = \mathbb{E}(Y^{Q_a}),$$

where $Q_a \sim \text{Categorical}\{q_a(A = 1 \mid V), \ldots, q_a(A = d \mid V)\}$.
These can adapt to positivity violations!

Issue: To respect positivity, may have non-zero probability of attending many treatments, and so $\psi_a - \psi_b$ can be driven by performance at treatment c. **Unfair?**

More generally: what does fairness mean?

Alternative: dynamic stochastic ("soft") interventions, but care is necessary

Dynamic stochastic intervention targeting treatment a:
For $b \in \{1, \ldots, d\}$, define an *interventional propensity score*

$$q_a(A = b \mid V) \quad \text{(a function of } V \subseteq X\text{)}.$$

The parameter targeting treatment a is

$$\psi_a = \mathbb{E}\big(Y^{Q_a}\big),$$

where $Q_a \sim \text{Categorical}\big\{q_a(A = 1 \mid V), \ldots, q_a(A = d \mid V)\big\}$.
These can adapt to positivity violations!

Issue: To respect positivity, may have non-zero probability of attending many treatments, and so $\psi_a - \psi_b$ can be driven by performance at treatment c. **Unfair?**

More generally: what does fairness mean?

Alternative: dynamic stochastic ("soft") interventions, but care is necessary

Dynamic stochastic intervention targeting treatment a:
For $b \in \{1, \ldots, d\}$, define an *interventional propensity score*

$$q_a(A = b \mid V) \quad \text{(a function of } V \subseteq X\text{)}.$$

The parameter targeting treatment a is

$$\psi_a = \mathbb{E}\left(Y^{Q_a}\right),$$

where $Q_a \sim \text{Categorical}\left\{q_a(A = 1 \mid V), \ldots, q_a(A = d \mid V)\right\}$.
These can adapt to positivity violations!

Issue: To respect positivity, may have non-zero probability of attending many treatments, and so $\psi_a - \psi_b$ can be driven by performance at treatment c. **Unfair?**

More generally: what does fairness mean?

Alternative: dynamic stochastic ("soft") interventions, but care is necessary

Dynamic stochastic intervention targeting treatment a:
For $b \in \{1, \ldots, d\}$, define an *interventional propensity score*

$$q_a(A = b \mid V) \quad \text{(a function of } V \subseteq X\text{)}.$$

The parameter targeting treatment a is

$$\psi_a = \mathbb{E}\left(Y^{Q_a}\right),$$

where $Q_a \sim \text{Categorical}\left\{q_a(A = 1 \mid V), \ldots, q_a(A = d \mid V)\right\}$.
These can adapt to positivity violations!

> Issue: To respect positivity, may have non-zero probability of attending many treatments, and so $\psi_a - \psi_b$ can be driven by performance at treatment c. **Unfair?**

More generally: what does fairness mean?

Fairness criterion

V-fairness criterion

In words, V-fairness: If a outperforms b in *every* stratum of V, then the parameters ψ_a and ψ_b must preserve that ordering (and same for $a = b$ and $a < b$)

For a subset $V \subseteq X$, parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are *V-fair* if

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \quad \Longrightarrow \quad \psi_a > \psi_b$$

and similarly for "$<$" and "$=$," under all allowed counterfactual distributions $\mathbb{P}^c$.

TSMs satisfy this property, but many parameters do not.

- Indirect standardization in provider profiling
- Naive implementations of dynamic stochastic interventions from causal inference

V-fairness criterion

In words, V-fairness: If a outperforms b in *every* stratum of V, then the parameters ψ_a and ψ_b must preserve that ordering (and same for $a = b$ and $a < b$)

For a subset $V \subseteq X$, parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are *V-fair* if

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \quad \Longrightarrow \quad \psi_a > \psi_b$$

and similarly for "$<$" and "$=$," under all allowed counterfactual distributions $\mathbb{P}^c$.

TSMs satisfy this property, but many parameters do not.

- Indirect standardization in provider profiling
- Naive implementations of dynamic stochastic interventions from causal inference

In words, V-fairness: If a outperforms b in *every* stratum of V, then the parameters ψ_a and ψ_b must preserve that ordering (and same for $a = b$ and $a < b$)

For a subset $V \subseteq X$, parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are *V-fair* if

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \quad \Longrightarrow \quad \psi_a > \psi_b$$

and similarly for "$<$" and "$=$," under all allowed counterfactual distributions $\mathbb{P}^c$.

TSMs satisfy this property, but many parameters do not.

- Indirect standardization in provider profiling
- Naive implementations of dynamic stochastic interventions from causal inference

V-fairness criterion

In words, V-fairness: If a outperforms b in *every* stratum of V, then the parameters ψ_a and ψ_b must preserve that ordering (and same for $a = b$ and $a < b$)

For a subset $V \subseteq X$, parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are *V-fair* if

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \quad \Longrightarrow \quad \psi_a > \psi_b$$

and similarly for "$<$" and "$=$," under all allowed counterfactual distributions $\mathbb{P}^c$.

TSMs satisfy this property, but many parameters do not.

- Indirect standardization in provider profiling
- Naive implementations of dynamic stochastic interventions from causal inference

In words, V-fairness: If a outperforms b in *every* stratum of V, then the parameters ψ_a and ψ_b must preserve that ordering (and same for $a = b$ and $a < b$)

> For a subset $V \subseteq X$, parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are *V-fair* if
>
> $$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \quad \implies \quad \psi_a > \psi_b$$
>
> and similarly for "$<$" and "$=$," under all allowed counterfactual distributions $\mathbb{P}^c$.

TSMs satisfy this property, but many parameters do not.

- Indirect standardization in provider profiling
- Naive implementations of dynamic stochastic interventions from causal inference

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \;=\; q_b(A = c \mid V) \quad \forall\, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness *if and only if* (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

Two intuitive properties $\iff$ V-fairness

How can we construct fair parameters?
Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \; = \; q_b(A = c \mid V) \quad \forall \, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness ***if and only if*** (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

Two intuitive properties $\iff$ V-fairness

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \;=\; q_b(A = c \mid V) \quad \forall\, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness *if and only if* (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

Two intuitive properties $\iff$ V-fairness

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) = q_b(A = c \mid V) \quad \forall c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^d$ facilitate pairwise V-fairness *if and only if* (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

Two intuitive properties $\iff$ V-fairness

How can we construct fair parameters?
Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \;=\; q_b(A = c \mid V) \quad \forall\, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness ***if and only if*** (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \; = \; q_b(A = c \mid V) \quad \forall\, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness ***if and only if*** (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \;=\; q_b(A = c \mid V) \quad \forall\, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness *if and only if* (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

How can we construct fair parameters?

Step 1: show its equivalence to two properties

Property 1: When targeting treatment a

- $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
  do not decrease target prop score
- $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
  Increase target prop. score for positive-prob set
- $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$
  do not increase non-target prop scores

Property 2: When comparing ψ_a and ψ_b

$$q_a(A = c \mid V) \; = \; q_b(A = c \mid V) \quad \forall \, c \notin \{a, b\}.$$

Make non-target prop scores equal

Lemma: $\{\psi_a\}_{a=1}^{d}$ facilitate pairwise V-fairness *if and only if* (Property 1 holds for all a) & (Property 2 holds for all a, b pairs).

Positivity

Minimum necessary positivity

What about positivity?

Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^d$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

Minimum necessary positivity

What about positivity?
Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \ \forall\, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^{d}$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

Minimum necessary positivity

What about positivity?
Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^d$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

Minimum necessary positivity

What about positivity?
Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^{d}$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

What about positivity?
Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^{d}$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

Minimum necessary positivity

What about positivity?
Some positivity is required.

Theorem: Let

$$C_V = \Big\{ v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \, a \Big\}.$$

denote the set of subjects that have non-zero prop score for every treatment.

Then, $\boxed{\mathbb{P}(C_V) > 0}$ is ***necessary*** for a set of fair parameters $\{\psi_a\}_{a=1}^{d}$ to be ***identifiable***

In other words, there has to be at least *some* subset of subjects with non-zero probability for each treatment.

Constructing fair and identifiable parameters

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \leq x$
Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b \neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b = a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}}$$
$$+ \underbrace{\mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \leq x$

Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b\neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b=a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}}$$
$$\underbrace{+ \mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \leq x$
Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b \neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b = a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}} \\ + \underbrace{\mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \leq x$
Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b \neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b=a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}} \underbrace{+ \mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \le x$

Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b \neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b = a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}}$$
$$+ \underbrace{\mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

Examples from properties 1 and 2 and min. positivity theorem

Properties 1 and 2: $f : [0,1] \to [0,1]$ such that $f(x) \leq x$

Any shift works; e.g., $f(x) = 0.9x$.

$$q_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a) f\{\pi_b(V)\}}_{\text{Shift prob. down at } b \neq a} + \underbrace{\mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]}_{\text{Shift prob. up at target } b=a}$$

1. Property 1 satisfied directly
2. Property 2: $q_a(A = c \mid V) = q_b(A = c \mid V) = f\{\pi_c(V)\}$

Necessary positivity: Require $\mathbb{P}(C_V) > 0$. Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}}$$
$$\underbrace{+ \mathbb{1}(V \in C_V)\left(\mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]\right)}_{\text{Intervention inside trimmed set}}$$

This construction may be of independent interest

1. This construction **generalizes** to any shift intervention by choosing $f(x)$; includes, e.g., incremental propensity score shift [Kennedy, 2019] and a risk ratio multiplication [Wen et al., 2023]

2. Generalizes dynamic stochastic interventions to un-ordered multi-valued treatment and **allows us to target specific treatments**.
   $\rightarrow$ **Future work:** target specific continuous treatment values

Identification and estimation

Identification & a doubly robust-style estimator

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Identification & a doubly robust-style estimator

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Identification & a doubly robust-style estimator

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Identification & a doubly robust-style estimator

Under consistency, exchangeability, and $\mathbb{P}(C_V) > 0$:

$$\psi_a = \mathbb{E}(Y^{Q_a}) = \mathbb{E}\Big[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\Big].$$

where $\mu_b(X) = \mathbb{E}(Y \mid A = b, X)$.

Issue 1: Plug-in estimators have bad properties.
Solution: Use doubly robust estimator

Issue 2: must estimate $\mathbb{1}(V \in C_V)$. This is difficult and precludes use of doubly robust or double ML estimators.
Solution: use smooth approximation $S(V \in C_V)$ for indicator.

Doubly robust-style estimator: $\widehat{\psi}_a = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_a(Z_i)$, where φ_a is *efficient influence function* of ψ_a.

Theorem: If $\sum_b \|\widehat{\pi}_b - \pi_b\| \big(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\big) = o_{\mathbb{P}}(n^{-1/2})$, then

$$\sqrt{n}(\widehat{\psi}_a - \psi_a) \rightsquigarrow N\big(0, \mathbb{V}\{\varphi_a(Z)\}\big)$$

Data analysis

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Data

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Data

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Data

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Data

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Setting: 10 largest dialysis providers in NY

- X: patient demographics, clinical characteristics, etc.
- A: dialysis provider
- Y: 30-day unplanned readmission (binary)

We constructed $q_a(A = b \mid X)$ to be X-fair but only require mild positivity on a *trimmed* subset of X.

We estimated the smooth trimmed TSMs across the 10 providers. These facilitate fair comparisons of all the providers.

Comparison vs. provider VIII

Provider VIII had the highest readmission rate

Provider	Diff. w.r.t. VIII	95% CI
I	0.019	[-0.006, 0.044]
II	0.027	[0.004, 0.050]
III	0.007	[-0.020, 0.033]
IV	0.012	[-0.013, 0.036]
V	0.009	[-0.022, 0.040]
VI	0.031	[0.002, 0.060]
VII	0.014	[-0.014, 0.043]
IX	0.013	[-0.019, 0.046]
X	0.029	[-0.003, 0.061]

Differences relative to VIII (highest readmission rate) show that II and VI appear significantly lower, while others are inconclusive.

Discussion

Recap:

1. Proposed a fairness criterion for comparing treatment efficacy
2. Established fairness criterion equivalent to two intuitive properties
3. Established minimum positivity condition for identification
4. Constructed fair and identifiable parameters
5. Constructed doubly robust-style efficient estimators

Recap:

1. Proposed a fairness criterion for comparing treatment efficacy
2. Established fairness criterion equivalent to two intuitive properties
3. Established minimum positivity condition for identification
4. Constructed fair and identifiable parameters
5. Constructed doubly robust-style efficient estimators

Recap:

1. Proposed a fairness criterion for comparing treatment efficacy
2. Established fairness criterion equivalent to two intuitive properties
3. Established minimum positivity condition for identification
4. Constructed fair and identifiable parameters
5. Constructed doubly robust-style efficient estimators

Recap:

1. Proposed a fairness criterion for comparing treatment efficacy
2. Established fairness criterion equivalent to two intuitive properties
3. Established minimum positivity condition for identification
4. Constructed fair and identifiable parameters
5. Constructed doubly robust-style efficient estimators

Recap:

1. Proposed a fairness criterion for comparing treatment efficacy
2. Established fairness criterion equivalent to two intuitive properties
3. Established minimum positivity condition for identification
4. Constructed fair and identifiable parameters
5. Constructed doubly robust-style efficient estimators

Thank you!

https://arxiv.org/abs/2410.13522

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Magdalena Bennett, Juan Pablo Vielma, and José R Zubizarreta. Building representative matched samples with multi-valued treatments in large observational studies. *Journal of computational and graphical statistics*, 29(4):744–757, 2020.

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Edward H Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, 2020.

Alexander W Levis, Edward H Kennedy, and Luke Keele. Nonparametric identification and efficient estimation of causal effects with instrumental variables. *arXiv preprint arXiv:2402.09332*, 2024.

Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13 (4):2389–2415, 2019.

Alexander R Luedtke and Mark J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Martin Roessler, Jochen Schmitt, and Olaf Schoffer. Can we trust the standardized mortality ratio? a formal analysis and evaluation based on axiomatic requirements. *PLoS One*, 16(9): e0257003, 2021.

Fredrik Sävje, Michael J Higgins, and Jasjeet S Sekhon. Generalized full matching. *Political Analysis*, 29(4):423–447, 2021.

Jeffrey H Silber, Paul R Rosenbaum, Richard N Ross, Justin M Ludwig, Wei Wang, Bijan A Niknam, Nabanita Mukherjee, Philip A Saynisch, Orit Even-Shoshan, Rachel R Kelz, et al. Template matching for auditing hospital cost and quality. *Health services research*, 49(5):1446–1474, 2014.

Jeffrey H Silber, Paul R Rosenbaum, Bijan A Niknam, Richard N Ross, Joseph G Reiter, Alexander S Hill, Lauren L Hochman, Sydney E Brown, Alexander F Arriaga, Rachel R Kelz, et al. Comparing outcomes and costs of surgical patients treated at major teaching and nonteaching hospitals: a national matched analysis. *Annals of surgery*, 271(3):412–421, 2020.

Lan Wen, Julia L Marcus, and Jessica G Young. Intervention treatment distributions that depend on the observed treatment process and model double robustness in causal survival analysis. *Statistical methods in medical research*, 32(3):509–523, 2023.

Backup

Criterion. (V-fairness) For a covariate subset $V \subseteq X$, the parameters $\psi_a = \mathbb{E}(Y^{Q_a})$ and $\psi_b = \mathbb{E}(Y^{Q_b})$ are V-fair if they satisfy:

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a > \psi_b, \tag{1}$$

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) = \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a = \psi_b, \text{ and} \tag{2}$$

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) < \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a < \psi_b \tag{3}$$

for all counterfactual distributions $\mathbb{P}^c$ in the counterfactual model.

- This is not an assumption. It is a **desideratum**; if we invoke the LHS, then the RHS holds.
- The relationship does **not** depend on $\mathbb{E}(Y^c \mid V)$ for $c \notin \{a, b\}$.
- The set of TSMs satisfy **V-fairness** for all $V \subseteq X$. If $\{\psi_a\}_{a=1}^d$ do also and LHS of (1), (2), or (3) hold, then $\{\psi_a\}_{a=1}^d$ have same ordering as TSMs.

V-fairness: contrapositives give further intuition

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a > \psi_b,$$
$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) = \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a = \psi_b, \text{ and}$$
$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) < \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a < \psi_b$$

$$\psi_a \leq \psi_b \implies \mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} < 1$$
$$\psi_a \neq \psi_b \implies \mathbb{P}^c\{\mathbb{E}(Y^a \mid V) = \mathbb{E}(Y^b \mid V)\} < 1$$
$$\psi_a \geq \psi_b \implies \mathbb{P}^c\{\mathbb{E}(Y^a \mid V) < \mathbb{E}(Y^b \mid V)\} < 1$$

V coarser $\rightarrow$ more desirable fairness condition
$\rightarrow$ e.g., $V = \emptyset$: $\psi_a \leq \psi_b \implies \mathbb{E}(Y^a) \leq \mathbb{E}(Y^b)$[1]

V more granular $\rightarrow$ less desirable fairness condition
$\rightarrow$ e.g., $V = X$: $\psi_a \leq \psi_b \implies \mathbb{P}^c\{\mathbb{E}(Y^a \mid X) > \mathbb{E}(Y^b \mid X)\} < 1$

[1]Combining three contrapositives yields equivalence:
$\psi_a < \psi_b \iff \mathbb{E}(Y^a) < \mathbb{E}(Y^b)$, etc.

V-fairness: other observations

$$\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} = 1 \implies \psi_a > \psi_b$$

- Roessler et al. [2021] provided **five axioms** for fairness. Our criterion is explicitly counterfactual and simpler (one condition).
- This is **not** algorithmic fairness.
- **Other antecedents**, like $\mathbb{P}(Y^a = Y^b) = 1$ or $\mathbb{P}(Y^a < y) < \mathbb{P}(Y^b < y)$ for all $y \in \mathbb{R}$, could be considered, but **conditional TSMs seemed most natural**
- We might want to **strengthen** the condition to **equivalence statement ($\iff$) for non-empty V**
  - Not possible because RHS involves **averages**, $\mathbb{E}(Y^{Q_a})$. Easy to construct counter-examples such that $\psi_a > \psi_b$ but $\mathbb{P}^c\{\mathbb{E}(Y^a \mid V) > \mathbb{E}(Y^b \mid V)\} < 1$.
  - However, with conditional curves, equivalence is possible! I.e., $\psi(\cdot) : \mathbb{P}^c \to (\mathcal{V} \to \mathbb{R})$ is a conditional map. *[Future work...]*

Two properties that are equivalent to fairness

~~Unintuitive fairness criterion~~ Intuitive properties

Let $\{q_a(A = b \mid V)\}_{b=1}^{d}$ denote the interventional propensity scores targeting treatment a.

Property 1:

- Do not decrease target propensity score
  $\mathbb{P}\{q_a(A = a \mid V) \geq \pi_a(V)\} = 1$
- Increase target prop. score for positive-prob set
  $\mathbb{P}\{q_a(A = a \mid V) > \pi_a(V)\} > 0$
- Do not increase non-target prop. scores
  $\mathbb{P}\{q_a(A = b \mid V) \leq \pi_b(V)\} = 1$ for all $b \neq a$

Let $\{q_a(A = c \mid V)\}_{c=1}^{d}$ and $\{q_b(A = c \mid V)\}_{c=1}^{d}$ denote interventional propensity scores target treatments a and b.

Property 2. Propensity scores equal at non-target treatments

$$\mathbb{P}\{q_a(A = c \mid V) = q_b(A = c \mid V)\} = 1 \text{ for all } c \notin \{a, b\}.$$

Lemma 1. Let $\{\psi_a\}_{a=1}^{d} = \{\mathbb{E}(Y^{Q_a})\}_{a=1}^{d}$ denote a set of parameters defined by dynamic stochastic interventions that vary with covariates $V \subseteq X$ and target treatments $1, \ldots, d$, respectively.

The set satisfies V-fairness for all $(a, b) \in \{1, \ldots, d\} \times \{1, \ldots, d\}$ if and only if it satisfies property 1 separately for all $a \in \{1, \ldots, d\}$ and property 2 for all $(a, b) \in \{1, \ldots, d\} \times \{1, \ldots, d\}$.

This is useful because **properties 1 and 2** are much more **intuitive**.

Properties 1 and 2 suggest examples

Property 1:

- Do not decrease target propensity score
- Increase target prop. score for positive-prob
- Do not increase non-target prop. scores

Property 2:

- Equal propensity scores for non-target treatments

For f satisfying $f(x) < x$,

$$\rho_a(A = b \mid V) = \underbrace{\mathbb{1}(b \neq a)}_{\text{non-target}} \underbrace{f\{\pi_b(V)\}}_{\text{decrease}} + \underbrace{\mathbb{1}(b = a)}_{\text{target}} \underbrace{\left[1 - \sum_{b \neq a} f\{\pi_b(V)\}\right]}_{\text{increase}}$$

For property 2, notice that

$$\rho_a(A = c \mid V) = \rho_b(A = c \mid V) = f\{\pi_c(V)\} \textit{ for } c \notin \{a, b\}$$

Properties 1 and 2 suggest examples

For f satisfying $f(x) < x$,

$$\rho_a(A = b \mid V) = \mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\left[1 - \sum_{b \neq a} f\{\pi_b(V)\}\right]$$

Examples of $f(\cdot)$:

- Multiplicative shift: $f(x) = \delta x, \delta < 1$,
- Exponential tilt: $f(x) = \frac{\delta x}{\delta x + 1 - x}, \delta < 1$,
- TSM: $f(x) = 0$.

By the way... This addresses complications from un-ordered treatment! **Intuition**: explicitly shift away from non-target treatments and implicitly towards target. This construction works for binary, multi-valued, and continuous treatment. *[Future work: target approx. dose-response curve]*

Minimum necessary positivity

Theorem 1. Let $\{\psi_a\}_{a=1}^d = \{\mathbb{E}(Y^{Q_a})\}_{a=1}^d$ and interventional propensity scores vary with V and let

$$C_V = \{v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \; a \in \{1, \ldots, d\}\} \quad (4)$$

denote the set of **subjects who have a non-zero probability of receiving every treatment**.

Then, $\mathbb{P}(C_V) > 0$ is necessary for the parameters to satisfy V-fairness and be identifiable simultaneously.

This does not depend on the parameters! It applies to any set of parameters that would satisfy V-fairness and be identifiable.

Trade-off: fairness versus positivity

Theorem 1 ($\mathbb{P}(C_V) > 0$) can also be stated as

$$\mathbb{E}\left(\prod_{a=1}^{d} \mathbb{1}\Big[\mathbb{P}\{\pi_a(X) > 0 \mid V\} = 1\Big]\right) > 0 \tag{5}$$

This framing can help illustrate trade-off between fairness and minimum positivity

$V = \emptyset$:

- Most desirable fairness condition;
- (5) $\equiv$ weak positivity (strong assumption)

$V = X$:

- Least desirable fairness condition
- (5) $\equiv \mathbb{E}\Big[\prod_{a=1}^{d} \mathbb{1}\{\pi_a(X) > 0\}\Big] > 0$. (weakest positivity assumption)

Identifiable examples and smooth approximations

Properties 1 & 2 plus minimum positivity suggest examples

Properties 1 and 2:

$$\rho_a(A = b \mid V) = \mathbb{1}(b \neq a) f\{\pi_b(V)\} + \mathbb{1}(b = a)\Big[1 - \sum_{b \neq a} f\{\pi_b(V)\}\Big]$$

Necessary positivity:

→ There must exist a set $C \in \mathcal{V}$ for which weak positivity holds.
→ "Trimmed" set:

$$C_V = \{v : \mathbb{P}\{\pi_a(X) > 0 \mid V = v\} = 1 \; \forall \; a \in \{1, \ldots, d\}\}$$

Only intervene on this set.

$$q_a(A = b \mid V) = \overbrace{\mathbb{1}(V \notin C_V)\pi_b(V)}^{\text{No intervention}} + \underbrace{\mathbb{1}(V \in C_V)\rho_a(A = b \mid V)}_{\text{Intervention inside trimmed set}}$$

Focusing on trimmed set agrees with prior intuition from matching and balancing

- In matching and balancing weights literature, prior work has emphasized focusing on group with non-zero probability of attending each treatment, arguing this facilitates **useful/fair comparisons between treatments** [Silber et al., 2014, 2020, Li and Li, 2019]

- Our work formalizes how/why this is fair

- Also suggests we could use **state-of-the-art matching methods** to construct $\mathbb{1}(V \in C_V)$ [Bennett et al., 2020, Sävje et al., 2021]

- We focus on **smooth approximation** of trimmed set instead

Smooth approximations of the trimmed set

$$q_a(A = b \mid V) = \mathbb{1}(V \notin C_V)\pi_b(V) + \mathbb{1}(V \in C_V)\rho_a(A = b \mid V)$$

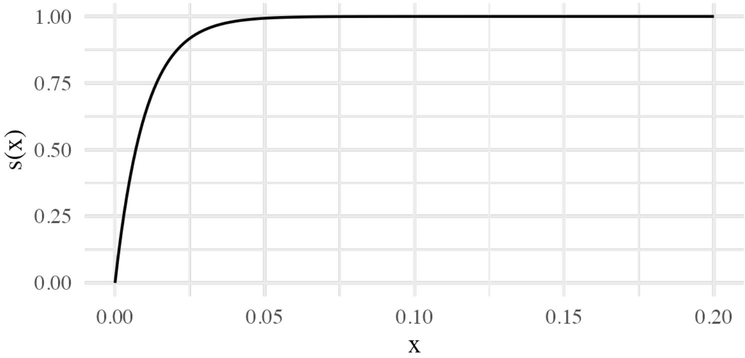
The indicator $\mathbb{1}(V \in C_V)$ is non-smooth, like in trimming
Solution: smooth approximation

$$\mathbb{1}(V \in C_V) = \mathbb{1}\left[\prod_{b=1}^{d} \mathbb{P}\left\{\pi_b(X) > 0 \mid V\right\} = 1\right] \approx \prod_{b=1}^{d} \mathbb{P}\left\{\pi_b(X) > 0 \mid V\right\}$$
$$= \prod_{b=1}^{d} \mathbb{E}\left[\mathbb{1}\left\{\pi_b(X) > 0\right\} \mid V\right]$$
$$\approx \prod_{b=1}^{d} \mathbb{E}\left[s\{\pi_b(X)\} \mid V\right] =: S(V \in C_V)$$

where $s\{\pi_b(X)\}$ approximates $\mathbb{1}\{\pi_b(X) > 0\}$.

$$q_a(A = b \mid V) = \{1 - S(V \in C_V)\}\pi_b(V) + S(V \in C_V)\rho_a(A = b \mid V)$$

Example: $s(x) = 1 - \exp(-kx), k = 100$



- Constraint that $s(0) = 0$ appears to be new (compared to prior trimming literature), and suggests novel smooth functions.
- It allows smooth parameters to be fair and identifiable
- We analyze generic $s(\cdot)$ satisfying $s(0) = 0$

Identification and estimation

Identification and plug-in estimator

Causal parameter: $\psi_a = \mathbb{E}\left(Y^{Q_a}\right)$ where

$$q_a(A = b \mid V) = \{1 - S(V \in C_V)\}\pi_b(V) + S(V \in C_V)\rho_a(A = b \mid V)$$

Suppose consistency, exchangeability, and the necessary positivity assumption ($\mathbb{P}(C_V) > 0$) hold. Then, **g-formula:**

$$\psi_a = \mathbb{E}\left[\sum_{b=1}^{d} \mathbb{E}\{\mu_b(X) \mid V\} q_a(A = b \mid V)\right]$$

Plug-in estimator can be biased or have slower-than-$\sqrt{n}$ convergence

Identification suggests the plug-in estimator

$$\widehat{\psi}_a = \mathbb{P}_n \left[\sum_{b=1}^{d} \widehat{\mathbb{E}}\{\widehat{\mu}_b(X) \mid V\} \widehat{q}_a(A = b \mid V) \right]$$

- $\mathbb{P}_n\{f(Z)\} = \frac{1}{n} \sum_{i=1}^{n} f(Z_i)$,
- $\widehat{\mu}_b(X)$ regresses $Y \sim \{X, \mathbb{1}(A = b)\}$,
- $\widehat{\mathbb{E}}\{\widehat{\mu}_b(X) \mid V\}$ regresses $\widehat{\mu}_b(X) \sim V$, and
- $\widehat{q}_a(A = b \mid V)$ plugs estimated propensity scores into the definition of $q_a(A = b \mid V)$

- With mis-specified parametric models, biased
- With nonparametric models, slower-than-$\sqrt{n}$ convergence

Theorem 2. Efficient influence function of ψ_a when $V = X$

$\psi_a = \mathbb{E}\left\{\sum_{b=1}^{d} \mu_b(X) q_a(A = b \mid X)\right\}$ and
$q_a(A = b \mid X) = \{1 - S(X \in C_X)\}\pi_b(X) + S(X \in C_X)\rho_a(A = b \mid X)$

$$\varphi_a(Z) = \sum_{b=1}^{d} \Bigg(\mu_b(X) q_a(A = b \mid X) + \left[\frac{\mathbb{1}(A = b)}{\pi_b(X)}\{Y - \mu_b(X)\}\right] q_a(A = b \mid X) + \mu_b(X)\varphi_{q_a}(Z; b) \Bigg)$$

where

$$\varphi_{q_a}(Z; b) = \varphi_S(Z)\{\rho_a(A = b \mid X) - \pi_b(X)\} + S(X \in C_X)\varphi_{\rho_a}(Z; b) + \Big\{1 - S(X \in C_X)\Big\}\Big\{\mathbb{1}(A = b) - \pi_b(X)\Big\},$$

$$\varphi_{\rho_a}(Z; b) = \mathbb{1}(b \neq a) f'\{\pi_b(X)\}\{\mathbb{1}(A = b) - \pi_b(X)\} - \mathbb{1}(b = a)\left[\sum_{b \neq a} f'\{\pi_b(X)\}\{\mathbb{1}(A = b) - \pi_b(X)\}\right], \text{ and}$$

$$\varphi_S(Z) = \sum_{b=1}^{d} \left(s'\{\pi_b(X)\}\{\mathbb{1}(A = b) - \pi_b(X)\} \right) \prod_{c \neq b}^{d} s\{\pi_c(X)\}.$$

Doubly robust-style estimator

Suppose $\{\widehat{\pi}_b, \widehat{\mu}_b\}_{b=1}^{d}$ constructed on independent sample.
Construct an estimator as

$$\widehat{\psi}_a = \mathbb{P}_n\{\widehat{\varphi}_a(Z)\}.$$

Can also use cross-fitting:

- Split data into K folds (5 or 10 common)
- Train $\widehat{\mu}$, $\widehat{\pi}$ on $K-1$ folds
- Evaluate on K^{th} fold.
- Cycle folds and repeat for full-sample efficiency

Theorem 3. Doubly robust-style estimator second order bias

The DR estimator bias satisfies (under some conditions)

$$\left|\mathbb{E}\left(\widehat{\psi}_a - \psi_a\right)\right| \lesssim$$

$$\sum_{b=1}^{d}\left|\mathbb{E}\left[\left\{\widehat{\pi}_b(X) - \pi_b(X)\right\}\left\{\widehat{\mu}_b(X) - \mu_b(X)\right\}\frac{\widehat{q}_a(A = b \mid X)}{\widehat{\pi}_b(X)}\right]\right|$$ Bias with q known

$$+\sum_{b=1}^{d}\|\widehat{\mu}_b - \mu_b\|\left[\sum_{c=1}^{d}\|\widehat{\pi}_c - \pi_c\|\right]$$ resid. prod. $\widehat{q}, \widehat{\mu}$

$$+\left\{\sum_{b=1}^{d}\|\widehat{\pi}_b - \pi_b\|\right\}\left\{\sum_{c=1}^{d}\|\widehat{\pi}_c - \pi_c\|\right\}$$ resid. prod. $\widehat{\rho}_a, \widehat{S}$

$$+\sum_{b\neq a}^{d}\|\widehat{\pi}_b - \pi_b\|^2$$ DR estimator ρ_a

$$+d\left\{\sum_{b=1}^{d}\|\widehat{\pi}_b - \pi_b\|^2 + \sum_{b=1}^{d}\sum_{c<b}\|\widehat{\pi}_b - \pi_b\|\,\|\widehat{\pi}_c - \pi_c\|\right\}$$ DR estimator S

What if we know the trimmed set?

When the trimmed set is known, then
$q_a(A = b \mid X) = \mathbb{1}(X \in C_X)\rho_a(A = b \mid X) + \mathbb{1}(X \notin C_X)\pi_b(X)$

$$\left|\mathbb{E}\left(\widehat{\psi}_a - \psi_a\right)\right| \lesssim$$

$$\sum_{b=1}^{d}\left|\mathbb{E}\left[\{\widehat{\pi}_b(X) - \pi_b(X)\}\{\widehat{\mu}_b(X) - \mu_b(X)\}\frac{\widehat{q}_a(A = b \mid X)}{\widehat{\pi}_b(X)}\right]\right| \quad \text{Bias with } q \text{ known}$$

$$+ \sum_{b=1}^{d}\|\widehat{\mu}_b - \mu_b\|\left[\mathbb{1}(b \neq a)\|f'(\pi_b)(\widehat{\pi}_b - \pi_b)\| + \mathbb{1}(b = a)\left\{\sum_{b \neq a}\|f'(\pi_b)(\widehat{\pi}_b - \pi_b)\|\right\}\right] \quad \text{resid. prod. } \widehat{\rho}, \widehat{\mu}$$

$$+ \sum_{b \neq a}^{d}\left\|f''(\pi_b)^{1/2}(\widehat{\pi}_b - \pi_b)\right\|^2 \quad \text{DR estimator } \rho_a$$

Similar type of bias term to more typical dynamic stochastic interventions (e.g., IPSIs)

What if we know the interventional propensity scores?

E.g., for trimmed TSMs, $f(\pi) = 0$ for all π.

$$\left|\mathbb{E}\left(\widehat{\psi}_a - \psi_a\right)\right| \lesssim$$

$$\sum_{b=1}^{d}\left|\mathbb{E}\left[\{\widehat{\pi}_b(X) - \pi_b(X)\}\{\widehat{\mu}_b(X) - \mu_b(X)\}\frac{\widehat{q}_a(A=b \mid X)}{\widehat{\pi}_b(X)}\right]\right| \quad \text{Bias with } q \text{ known}$$

$$+ \sum_{b=1}^{d}\left(\|\widehat{\mu}_b - \mu_b\| + \|\widehat{\pi}_b - \pi_b\|\right)\left\{\sum_{c=1}^{d}\|\widehat{\pi}_c - \pi_c\|\right\} \quad \text{resid. prod. } \widehat{q}, \widehat{\mu}$$

$$+ d\left\{\sum_{b=1}^{d}\|\widehat{\pi}_b - \pi_b\|^2 + \sum_{b=1}^{d}\sum_{c<b}\|\widehat{\pi}_b - \pi_b\|\,\|\widehat{\pi}_c - \pi_c\|\right\} \quad \text{DR estimator } S$$

$S(X \in C_X) = \prod_{b=1}^{d} s\{\pi_b(X)\}$ **involves products of propensity scores**: yields the double sums and outer factor d.

What if we know the trimmed set and the interventional propensity scores?

If both the trimmed set and $f(\pi)$ were known, then

- the estimand simplifies to

$$\psi_a = \mathbb{E}\{\mu_a(X)\mathbb{1}(X \in C_X)\} + \mathbb{E}\{Y\mathbb{1}(X \notin C_X)\}$$

- the doubly robust-style estimator is

$$\widehat{\psi}_a := \mathbb{P}_n\left(\left[\frac{\mathbb{1}(A = a)}{\widehat{\pi}_a(X)}\{Y - \widehat{\mu}_a(X)\} + \widehat{\mu}_a(X)\right]\mathbb{1}(X \in C_X) + Y\mathbb{1}(X \notin C_X)\right)$$

Canonically doubly robust bias (over C_X):

$$\left|\mathbb{E}\left(\widehat{\psi}_a - \psi_a\right)\right| \lesssim \mathbb{E}\left[\{\widehat{\pi}_a(X) - \pi_a(X)\}\{\widehat{\mu}_a(X) - \mu_a(X)\}\mathbb{1}\left(X \in C_X\right)\right]$$

Normal limiting distribution

Under conditions of main theorem, suppose also that $\sum_{a=1}^{d} \mathbb{E}\{\widehat{\varphi}_a(Z) - \varphi_a(Z)\} = o_{\mathbb{P}}(1)$ and

$$\sum_{a=1}^{d}\sum_{b=1}^{d} \|\widehat{\mu}_a - \mu_a\| \|\widehat{\pi}_b - \pi_b\| + d \sum_{a=1}^{d}\sum_{b \leq a} \|\widehat{\pi}_a - \pi_a\| \|\widehat{\pi}_b - \pi_b\| = o_{\mathbb{P}}(n^{-1/2}).$$

Then,

$$\sqrt{n} \begin{pmatrix} \widehat{\psi}_1 - \psi_1 \\ \vdots \\ \widehat{\psi}_d - \psi_d \end{pmatrix} \rightsquigarrow N(0, \Sigma)$$

where $e_i^T \Sigma e_j = \text{cov}\{\varphi_i(Z), \varphi_j(Z)\}$.

- $\sqrt{n}$ convergence to Gaussian possible under $n^{-1/4}$ rate conditions on nuisance estimators
- Many typical analyses in provider profiling follow by the delta method

Data analysis

10 largest providers (by number of claims) in NY state

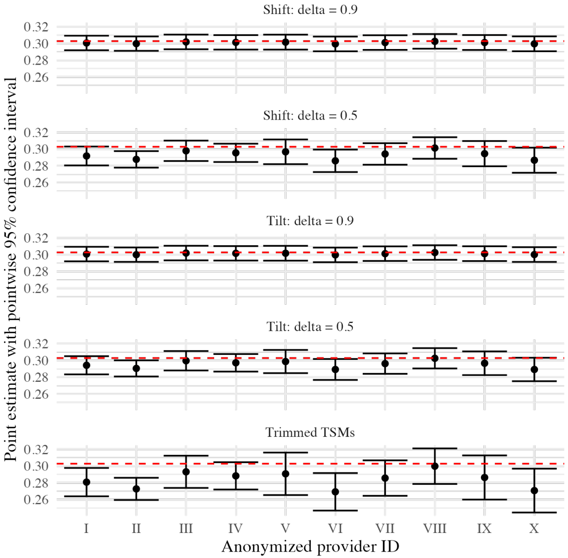
- X: physical attributes, social factors, clinical characteristics, demographic info
- A: the provider attended
- Y: 30-day unplanned readmission (binary)

We considered parameters that are X-fair (so, weak fairness condition) but require only mild positivity assumption for ID.

Constructed $S(X \in C_V)$ using $s(x) = 1 - \exp(-100x)$.

Five interventions: multiplicative shifts and exponential tilts with $\delta \in \{0.5, 0.9\}$, and trimmed TSMs.

Point estimates and 95% CIs



What about compared to provider VIII? (which had the worst/highest readmission rate)

Anonymized provider identifier	Difference between 30-day readmission rate for provider VIII and this provider	95% confidence interval
I	0.019	[-0.006, 0.044]
II	0.027	[0.004, 0.050]
III	0.007	[-0.020, 0.033]
IV	0.012	[-0.013, 0.036]
V	0.009	[-0.022, 0.040]
VI	0.031	[0.002, 0.060]
VII	0.014	[-0.014, 0.043]
IX	0.013	[-0.019, 0.046]
X	0.029	[-0.003, 0.061]

Future work

Future work

- We formalized the intuition that **it's only fair to compare treatments when subjects could reasonably take both**. Therefore, methods for partitioning the covariate/treatment space could be useful.
  - Intuition suggests, e.g., partition along state lines. Perhaps **data-adaptive** approach could be better.
- Dependence on $\widehat{\pi}$ and d is disappointing, but makes sense given the construction of $S(X \in C_X)$.
  - Next step: figure out how to **dodge dimension dependence.** Positivity violations suggest there is a **sparsity phenomenon**, so may be able to construct better estimators leveraging that.
  - Alternatively/additionally, consider unfair interventions that are easier to estimate, and **quantify how unfair** they are.
- **Heterogeneity:** inherently interesting, and could satisfy stronger/different fairness condition.

If $s(x) = 1 - \exp(-k_n x)$ where k_n can vary with sample size,

$$\left|\mathbb{E}\left(\widehat{\psi}_a - \psi_a\right)\right| \lesssim \sum_{b=1}^{d}\sum_{c=1}^{d} k_n \|\widehat{\mu}_b - \mu_b\| \|\widehat{\pi}_c - \pi_c\| + dk_n^2 \left(\|\widehat{\pi}_b - \pi_b\| \|\widehat{\pi}_c - \pi_c\|\right).$$

Consider $\mathbb{E}(Y^Q \mid A = a)$ where
$Q \sim \text{Categorical}\{q(A = 1 \mid X), \ldots, q(A = d \mid X)\}$. Then,

$$
\begin{aligned}
\mathbb{E}(Y^Q \mid A = a) &= \mathbb{E}\left\{\mathbb{E}(Y^Q \mid A = a, X) \mid A = a\right\} \\
&= \mathbb{E}\left\{\sum_b \mathbb{E}(Y^b \mid A = a, X) q(A = b \mid X) \mid A = a\right\} \\
&= \int_{\mathcal{X}} \sum_b \mathbb{E}(Y^b \mid X = x) q(A = b \mid X = x) d\mathbb{P}(X = x \mid A = a) \\
&= \int_{\mathcal{X}} \sum_b \mathbb{E}(Y^b \mid X = x) q(A = b \mid X = x) \frac{\pi_a(X = x)}{\mathbb{P}(A = a)} d\mathbb{P}(x) \\
&= \mathbb{E}\left\{\mathbb{E}(Y^b \mid X) \frac{q(A = b \mid X)\pi_a(X)}{\mathbb{P}(A = a)}\right\}
\end{aligned}
$$

$$
= \mathbb{E}(Y^{Q'}), Q' \sim \text{Categorical}\left\{\frac{q(A = 1 \mid X)\pi_a(X)}{\mathbb{P}(A = a)} ..., \frac{q(A = d \mid X)\pi_a(X)}{\mathbb{P}(A = a)}\right\}
$$

by (1) IE, (2) definition of intervention, (3) exchangeability, (4) Bayes', (5) definition of expectation, (6) re-defining intervention.

Margin conditions

Instead of smooth approximations, can use margin conditions [Levis et al., 2024, Kennedy et al., 2020, Audibert and Tsybakov, 2007, Luedtke and van der Laan, 2016].

$\alpha > 0$ such that for all $a \in \{1, \ldots, d\}$ and any $t \geq 0$,

$$\mathbb{P}\{|\pi_a(X) - 0| \leq t\} \lesssim t^\alpha$$

However, this is not useful because it enforces $\mathbb{P}\{\pi_a(X) = 0\} = 0$ and, when applied across all $a \in \{1, \ldots, d\}$, imposes weak positivity.