Assessing Disparate Impacts of Personalized Interventions: Identifiability and Bounds

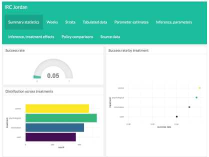
"What works for whom"

Interventions and impact evaluation enact "data-driven decision-making" in the public sector and social services.

- Some criticism of fair ml: suggesting interventions over predictions, e.g. Barabas et al. [2017]
- Nonetheless, interventions under scarcity will be triaged based on predictions
- "True labels" are **not observed in interventional data** because of the "fundamental problem of causal inference"
- Interest in personalizing interventions based on predictions dating back to Behncke et al. [2007]: randomized caseworkers to predictions of job training program effects

Related work: Interventions and Fairness

- Methodology for predicting "what works": CATE estimation: Shalit et al. [2017], Wager and Athey [2017]
- Personalized allocations: Kube and Das [2019] homelessness services
- Quinn et al. [2019]: MAB trial for job training to refugees in Jordan



Personalized interventions

Data (RCT or observational), (X, A, T, Y), on individuals:

- Prognostic features X (for personalization)
- Sensitive attribute A
- Binary treatment indicator $T \in \{0, 1\}$
- Binary response outcome $Y(0), Y(1) \in \{0, 1\}$ (benefit to the individual)
- Fundamental problem of causal inference: we only observe $Y(T)$
- Decision rules, $Z(X, A) \in \{0, 1\}$

Personalized Allocations

Conditional average treatment effect τ

$$\begin{aligned}\tau = \tau(X, A) &= \mathbb{E}[Y(1) - Y(0) \mid X, A] \\ &= \mathbb{P}(Y = 1 \mid T = 1, X, A) - \mathbb{P}(Y = 1 \mid T = 0, X, A),\end{aligned}$$

We assess disparities induced by Z, e.g. thresholds on τ,

$$Z = \mathbb{I}[\tau < \theta]$$

True positive rate: responders

Enumerate all possible potential outcomes:

- $Y(0) < Y(1)$: responders
- $Y(0) > Y(1)$: anti-responders
- $Y(0) = Y(1)$: non-responders.

True positive rate for interventions:
of those *who would actually benefit from treatment*,
how many were allocated treatment under Z?

$$\mathrm{TPR}_a = \mathbb{P}(Z = 1 \mid A = a, Y(1) > Y(0)),$$
$$\mathrm{TNR}_a = \mathbb{P}(Z = 0 \mid A = a, Y(1) \leq Y(0)).$$

Fairness for Personalized interventions

Denote

$$p_{yy'}(x) : \mathbb{P}(Y(0) = y, Y(1) = y' \mid X = x)$$

We want to estimate $p_{01} = \mathbb{P}(Y(1) > Y(0) \mid X)$,
but crucially we can only estimate:

$$\mathbb{P}(Y \mid T = 1, X) = p_{11} + p_{01}$$
$$\mathbb{P}(Y \mid T = 0, X) = p_{10} + p_{11}$$

Proposition 1: TPR, TNR are **not identifiable** from the data.

Fairness for Personalized interventions

Identification Assumption 1: Treatment monotonicity.

$$\boxed{Y(1) \geq Y(0)}$$

i.e., $\boxed{p_{10}} = 0$: Job training cannot *hurt* one's ability to get a job.

Proposition 2: (Identification under monotonicity).

$$\mathrm{TPR}_a = \frac{\mathbb{E}\left[\tau \mid A = a, Z = 1\right] \mathbb{P}\left(Z = 1 \mid A = a\right)}{\mathbb{E}\left[\tau \mid A = a\right]},$$

$$\mathrm{TNR}_a = \frac{\mathbb{E}\left[(1-\tau) \mid A = a, Z = 0\right] \mathbb{P}\left(Z = 0 \mid A = a\right)}{\mathbb{E}\left[(1-\tau) \mid A = a\right]}.$$

Sensitivity Analysis

Assumption 2: B-relaxed monotone treatment response:

$$\boxed{p_{10} \leq B}$$

Anti-responder probability is uniformly less than B.

$$\rho_a^{\mathrm{TPR}}(\boxed{\eta}) := \frac{\mathbb{E}\left[\tau + \boxed{\eta} \mid A = a, Z = 1\right] \mathbb{P}\left(Z = 1 \mid A = a\right)}{\mathbb{E}\left[\tau + \boxed{\eta} \mid A = a\right]}$$

Given an uncertainty set $\mathcal{U}$ for p_{10}, we define the simultaneous identification region of the TPR and TNR for all groups $a \in \mathcal{A}$ as:

$$\boxed{\Theta} = \{\left(\rho_a^{\mathrm{TPR}}(\eta), \rho_a^{\mathrm{TNR}}(\eta)\right)_{a \in \mathcal{A}} : \eta \in \mathcal{U}\} \subseteq \mathbb{R}^{2 \times |\mathcal{A}|}.$$

Proposition 4: Support function access to the convex hull of the identified set.
Let $r_a^z := \mathbb{P}(Z = z \mid A = a)$ and $\tau_a^z := \mathbb{E}[\tau \mid A = a, Z = z]$. For sets $\mathcal{U}$ which are product sets over groups,

$$h_\Theta(\mu) := \sup_{\rho \in \Theta} \mu^\top \rho, \tag{1}$$

Eq. (1) can be reformulated as:

$$\begin{aligned} h_\Theta(\mu) = & \sum_{a \in \mathcal{A}} h_{\Theta_a}(\mu_a) \\ h_{\Theta_a}(\mu_a) = & \sup_{\omega_a, t_a} \quad \begin{array}{l} \mu_a^{\mathrm{TPR}} r_a^1 \left(t_a \tau_a^1 + \mathbb{E}\left[\omega_a(X) \mid {A=a \atop Z=1}\right]\right) + \\ \frac{\mu_a^{\mathrm{TNR}} r_a^0}{t_a - 1} \left(t_a \left(1 - \tau_a^0\right) + \mathbb{E}\left[\omega_a(X) \mid {A=a \atop Z=0}\right]\right) \end{array} \\ & \quad \text{s.t.} \quad \omega_a(\cdot) \in t_a\, \mathcal{U}_a, \quad t_a \left(r_a^0 \tau_a^0 + r_a^1 \tau_a^1 \right) + \mathbb{E}\left[\omega_a \mid A = a\right] = 1. \end{aligned}$$

Assessment on clinical data

Proposition 5 (informal): Under Asn. 2, the sharp identification intervals for TPR_a and TNR_a (each) are closed-form intervals, respectively. Moreover, these intervals are tight;

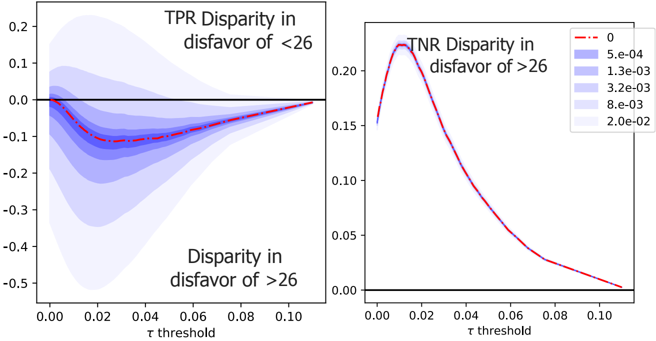
$$(\underline{\rho}_a^{\mathrm{TPR}}(B), \underline{\rho}_a^{\mathrm{TNR}}(B)) \in \Theta_{B,a}$$
$$(\overline{\rho}_a^{\mathrm{TPR}}(B), \overline{\rho}_a^{\mathrm{TNR}}(B)) \in \Theta_{B,a},$$

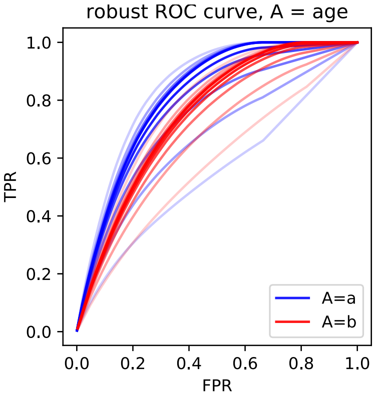
i.e., the two extremes are simultaneously achievable.

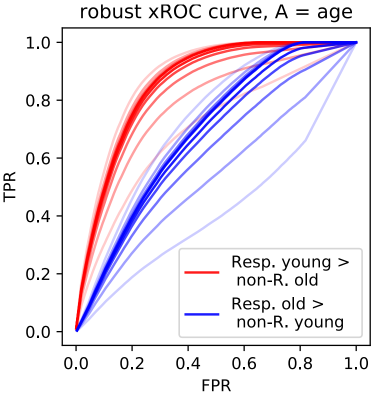
Job training case study

- Behaghel et al. [2014]: French job training program; n = 11k. Train τ using generalized random forest [Wager and Athey, 2017].
- $T = 1$ is assignment to public job training program;
  $T = 0$ is control.
  ATE $= 0.02$ (significant)
- $Y = 1$ is employment at 6 months, $A = 1$ is age > 26.

Job training case study




robust ROC curve, A = age
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
A=a
A=b


robust xROC curve, A = age
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Resp. young > non-R. old
Resp. old > non-R. young