Debiased Front-Door Learners For Heterogeneous Effects

Yonghan Jung
University of Illinois Urbana-Champaign
School of Information Sciences

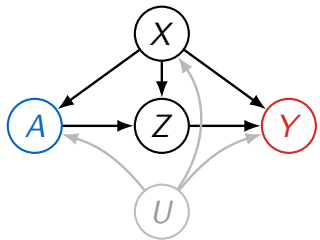
ACIC 2026 Oral
ICLR 2026
May 13, 2026

Paper
arXiv:2509.22531

Code
GitHub

Front-Door Model

Causal Graph



Variables

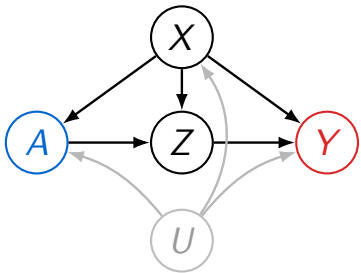
- X observed covariates
- **A** treatment
- Z mediator
- **Y** outcome
- **U** unmeasured confounder

Front-Door Use Case

- Treatment and outcome may be confounded.
- A mediator pathway $A \to Z \to Y$ is observed.
- A mediator is not affected by unmeasured confounders.

References: Pearl (1995); Fulcher et al. (2020).

Front-Door Model In Real-World Studies



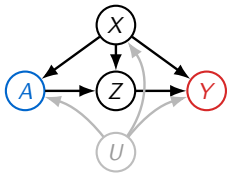
2024 Rothman Epidemiology Prize

Piccininni et al., Epidemiology 2023

- **A** mobile stroke unit care
- **Z** time to thrombolysis
- **Y** cardiovascular functional outcome
- **X** measured clinical and system covariates
- **U** unmeasured care-selection factors

Sources: Epidemiology 35(4):431, 2024; Epidemiology 34(5):712-720, 2023.

Front-Door Identification



Graphical FD Criterion

FD1 all directed $A \to Y$ paths pass through Z
FD2 all back-door A–Z paths are blocked by X
FD3 all back-door Z–Y paths are blocked by (A, X)
Pos. $p(a \mid x) > 0$ and $p(z \mid a, x) > 0$

Identification

$$\begin{aligned}\tau_{\bar{a}}(x) &\triangleq \mathbb{E}[Y(\bar{a}) \mid X = x] \\ &= \sum_z \Pr(Z = z \mid A = \bar{a}, X = x) \sum_{a'} \Pr(A = a' \mid X = x)\mathbb{E}[Y \mid Z = z, A = a', X = x], \\ \tau(x) &\triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau_1(x) - \tau_0(x).\end{aligned}$$

References: Pearl (1995); Fulcher et al. (2020).

Position of This Work

Existing literature

- Front-door identification
  Pearl (1995)
- Robust / efficient FD estimation
  Fulcher et al. (2020); Guo et al. (2023); Wen et al. (2024)
- Conditional FD estimation using ML
  Xu and Gretton (2022); Xu et al. (2024); Chen et al. (2025)
- Robust HTE learners (under ignorability)
  Chernozhukov et al. (2018); Nie and Wager (2021); Kennedy (2023); Foster and Syrgkanis (2023)

Position

front-door model
+
HTE learning
+
orthogonal to nuisance estimates
(NOT first-order dependent)

This work:
We develop two robust CATE estimators under front-door structure.

Nuisances

$$
\begin{aligned}
m_{za}(x) &\triangleq \mathbb{E}[Y \mid Z = z, A = a, X = x] && \text{outcome response,} \\
e_a(x) &\triangleq \Pr(A = a \mid X = x) && \text{treatment response,} \\
q_{za}(x) &\triangleq \Pr(Z = z \mid A = a, X = x) && \text{mediator response.}
\end{aligned}
$$

$$
\tau_{\bar{a}}(x) \triangleq \mathbb{E}[Y(\bar{a}) \mid X = x] = \sum_z q_{z\bar{a}}(x) \sum_{a'} e_{a'}(x) m_{za'}(x),
$$

$$
\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \sum_z \{q_{z1}(x) - q_{z0}(x)\} \sum_{a'} e_{a'}(x) m_{za'}(x).
$$

When Z and A are binary,

$$
\tau(x) = \{q_{11}(x) - q_{10}(x)\} \sum_{a' \in \{0,1\}} e_{a'}(x)\{m_{1a'}(x) - m_{0a'}(x)\}.
$$

Plug-In Inherits First-Order Nuisance Error

Plug-In Estimator: estimate the nuisance functions (m, e, q), then plug $(\hat{m}, \hat{e}, \hat{q})$ into the front-door formula from the previous slide.

$$\begin{aligned}\hat{\tau}_{\mathrm{PI}}(x) &\triangleq T(\hat{m}, \hat{e}, \hat{q})(x) \\ &= \sum_{z} \{\hat{q}_{z1}(x) - \hat{q}_{z0}(x)\} \sum_{a'} \hat{e}_{a'}(x) \hat{m}_{za'}(x).\end{aligned}$$

The first-order expansion is

$$T(\hat{m}, \hat{e}, \hat{q}) - T(m, e, q) \approx D_m T[\hat{m} - m] + D_e T[\hat{e} - e] + D_q T[\hat{q} - q].$$

$D_m T, D_e T, D_q T$ are derivative maps: nuisance error enters the target at first order.

Overview

Target: $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

We develop two robust conditional estimators under front-door structure.

Learner	Template	Core construction
FD-DR	DR-learner	FD pseudo-outcome; regress on X
FD-R	R-learner	FD pathway decomposition; R-loss stage fits

DR-learner: Kennedy (2023); R-learner: Nie and Wager (2021).

FD-DR: Pseudo-Outcome Target

Let $V \triangleq (Y, Z, A, X)$ and $\eta_0 \triangleq (m, e, q)$ be the true FD nuisance set.

Object	**Identity**
FD pseudo-outcome	$\varphi_{\bar{a}}(V; \eta)$
CATE recovery	$\mathbb{E}[\varphi_{\bar{a}}(V; \eta_0) \mid X = x] = \tau_{\bar{a}}(x)$
ATE recovery	$\mathbb{E}[\varphi_{\bar{a}}(V; \eta_0)] = \mathbb{E}[\tau_{\bar{a}}(X)]$
CATE label	$D_{\eta_0}(V) \triangleq \varphi_1(V; \eta_0) - \varphi_0(V; \eta_0)$, $\mathbb{E}[D_{\eta_0}(V) \mid X = x] = \tau(x)$.

FD-DR builds a debiased label whose conditional mean is the front-door CATE.

FD-DR: Front-Door Pseudo-Outcome

$$\varphi_{\bar{a}}(V;\eta) \triangleq \underbrace{\xi_{\bar{a}}(V)\{Y - m_{ZA}(X)\}}_{\text{outcome residual}} + \underbrace{\pi_{\bar{a}}(V)\{r(Z,X) - \nu(A,X)\}}_{\text{treatment residual correction}} + \underbrace{s_{\bar{a}}(A,X)}_{\text{FD plug-in pathway}},$$

$$\xi_{\bar{a}}(V) \triangleq \frac{q_{Z\bar{a}}(X)}{q_{ZA}(X)}, \qquad \pi_{\bar{a}}(V) \triangleq \frac{\mathbf{1}\{A = \bar{a}\}}{e_{\bar{a}}(X)},$$

$$r(z,X) \triangleq \sum_{a'} m_{za'}(X) e_{a'}(X), \qquad \nu(a,X) \triangleq \sum_{z} r(z,X) q_{za}(X), \qquad s_{\bar{a}}(a,X) \triangleq \sum_{z} m_{za}(X) q_{z\bar{a}}(X).$$

FD-DR Algorithm

1. **Two-way split.** Use I_η for nuisance learning and I_τ for the final CATE regression.
2. **Fit FD nuisances.** On I_η, estimate $\hat{\eta} = (\hat{m}, \hat{e}, \hat{q})$.
3. **Evaluate FD-POs.** For $i \in I_\tau$,

$$\hat{D}_i \triangleq \varphi_1(V_i; \hat{\eta}) - \varphi_0(V_i; \hat{\eta}).$$

4. **Regress labels on covariates.**

$$\hat{\tau}_{\mathrm{DR}} \in \arg\min_{\tau \in \mathcal{T}} \sum_{i \in I_\tau} \{\hat{D}_i - \tau(X_i)\}^2.$$

Use K-fold cross-fitting in practice by rotating the two roles and pooling the out-of-fold DR labels.

FD-DR Theorem: Error Rate

Define

$$\delta_m \triangleq \|\hat{m} - m\|_2,$$
$$\delta_e \triangleq \max_a \|\hat{e}_a - e_a\|_2,$$
$$\delta_q \triangleq \max_{z,\bar{a}} \|\hat{q}_{z\bar{a}} - q_{z\bar{a}}\|_2.$$

With R_{DR} denoting final-stage regresion error, under bounded overlap,

$$\|\hat{\tau}_{\mathrm{DR}} - \tau\|_2^2 = O_p\big(R_{\mathrm{DR}} + \delta_q^2(\delta_m^2 + \delta_e^2)\big)\,.$$

If $\delta_m, \delta_e, \delta_q = O_p(n^{-1/4})$,

$$\|\hat{\tau}_{\mathrm{DR}} - \tau\|_2^2 = O_p\big(n^{-1} + R_{\mathrm{DR}}\big)\,.$$

FD-R: Brief Idea

$$\textbf{Effect } (A \to Y) = (A \to Z) \textbf{ link} \times \textbf{averaged } (Z \to Y) \textbf{ link}$$

$$\tau(x) = \underbrace{\{q_{11}(x) - q_{10}(x)\}}_{A \to Z} \underbrace{\sum_{a \in \{0,1\}} e_a(x)\{m_{1a}(x) - m_{0a}(x)\}}_{\gamma_g(x):\ Z \to Y,\ \text{averaged over } A}.$$

Define

$$b(x) \triangleq q_{11}(x) - q_{10}(x) \qquad (A \to Z),$$

$$g(a,x) \triangleq m_{1a}(x) - m_{0a}(x) \qquad (Z \to Y),$$

$$\gamma_g(x) \triangleq \sum_{a \in \{0,1\}} e_a(x) g(a,x) = e_0(x) g(0,x) + e_1(x) g(1,x) \qquad \text{averaged } (Z \to Y).$$

Then,

$$\tau(x) = b(x)\gamma_g(x).$$

FD-R: Robinson Decomposition

$b(x)$ and $g(a, x)$ can be estimated using R-learners (Nie and Wager (2021)).

For the $A \to Z$ link:

$$Z - \mathbb{E}[Z \mid X] = \{A - \mathbb{E}[A \mid X]\} b(X) + \varepsilon_Z.$$

For the $Z \to Y$ link:

$$Y - \mathbb{E}[Y \mid A, X] = \{Z - \mathbb{E}[Z \mid A, X]\} g(A, X) + \varepsilon_Y.$$

FD-R: Plug-In Error For γ_g

To estimate

$$\gamma_g(x) = e_0(x)g(0,x) + e_1(x)g(1,x),$$

we consider

$$\hat{\gamma}_g^{\text{plug}}(x) \triangleq \hat{e}_{A=0}(x)\hat{g}(0,x) + \hat{e}_{A=1}(x)\hat{g}(1,x).$$

Then

$$\hat{\gamma}_g^{\text{plug}}(x) - \gamma_g(x) \approx \Delta_g(x) + \{\hat{e}_{A=1}(x) - e_{A=1}(x)\}\{g(1,x) - g(0,x)\}.$$

where

$$\Delta_g(x) \triangleq \mathbb{E}[\hat{g}(A,x) - g(A,x) \mid X = x].$$

In other words, $\hat{\gamma}_g^{\text{plug}}$ is first-order dependent on the nuisance $\hat{e}$.

FD-R: Estimating γ_g By Error Correction

Build the corrected pseudo-g label

$$\hat{\zeta}_g(A, X) \triangleq \hat{\gamma}_g^{\text{plug}}(X) + \underbrace{\{A - \hat{e}_{A=1}(X)\} \times \{\hat{g}(1, X) - \hat{g}(0, X)\}}_{\text{error correction}}.$$

Its conditional bias is

$$\mathbb{E}[\hat{\zeta}_g(A, X) \mid X = x] - \gamma_g(x) = \Delta_g(x).$$

The remaining bias no longer has first-order dependence on the propensity plug-in error.

FD-R Algorithm

0. **Split dataset into I_η, I_{bg}, I_γ.**
1. **I_η: fit nuisance functions.**

$$\hat{\eta}_b = (\hat{e}_A, \hat{m}_Z), \qquad \hat{\eta}_g = (\hat{e}_Z, \hat{m}_Y).$$

$e_A(x) = \mathbb{E}[A \mid X = x]$, $m_Z(x) = \mathbb{E}[Z \mid X = x]$; $e_Z(a, x) = \mathbb{E}[Z \mid A = a, X = x]$, $m_Y(a, x) = \mathbb{E}[Y \mid A = a, X = x]$.

2. **I_{bg}: fit two R-learner stages.**

$$A \to Z: \ \hat{b}(X), \qquad Z \to Y: \ \hat{g}(A, X).$$

3. **I_γ: regress corrected pseudo-g labels and return.**

$$\hat{\zeta}_g(A, X) \ \rightsquigarrow \ \hat{\gamma}_g(X), \qquad \hat{\tau}_{\mathrm{R}}(X) \triangleq \hat{b}(X)\hat{\gamma}_g(X).$$

FD-R Theorem: Error Rate

$$\mathcal{E}_{\mathrm{R}} \triangleq R_b + R_g + R_\gamma \qquad \text{(Regression-stage errors)}.$$

Stable regression assumption over $\hat{\gamma}$:

$$\|\hat{\gamma}_{\hat{g}} - \bar{\gamma}_{\bar{g}}\|_{2,C}^2 \lesssim \|\hat{g} - \bar{g}\|_{2,XC}^2.$$

$$\|\hat{\tau}_{\mathrm{R}} - \tau\|_2^2 = O_p\left(\mathcal{E}_{\mathrm{R}} + \delta_{e_A}^4 + \delta_{m_Z}^2 \delta_{e_A}^2 + \delta_{e_Z}^4 + \delta_{m_Y}^2 \delta_{e_Z}^2\right).$$

If all four nuisance rates are $O_p(n^{-1/4})$, then

$$\|\hat{\tau}_{\mathrm{R}} - \tau\|_2^2 = O_p\left(\mathcal{E}_{\mathrm{R}} + n^{-1}\right).$$

FD-DR Vs FD-R

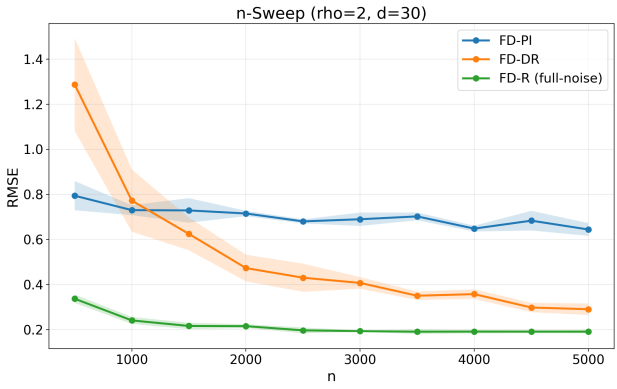
FD-DR

- \+ Product-error debiased FDPO for $\tau(X)$.
- \+ Attractive when FDPO products are small and overlap is credible.
- — Uses inverse weights and density ratios (π, ξ).
- — Variance can inflate when e or q approach 0 or 1.

FD-R

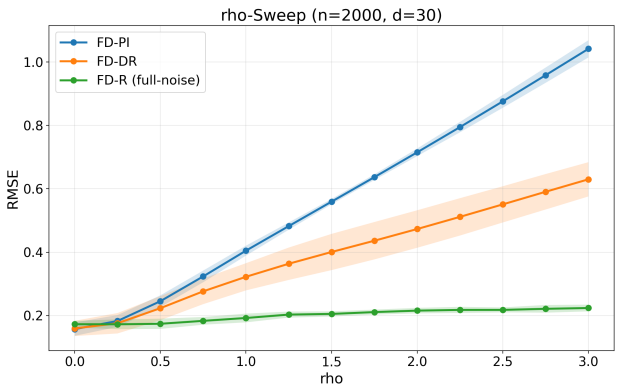
- \+ Avoids the density ratios required by FD-DR.
- \+ Gives pathway diagnostics through $b(X)$, $g(A, X)$, and $\gamma_g(X)$.
- — Non-doubly robustness.
- — Costs extra nuisance and stage fits.

Synthetic Results: Sample Size Sweep



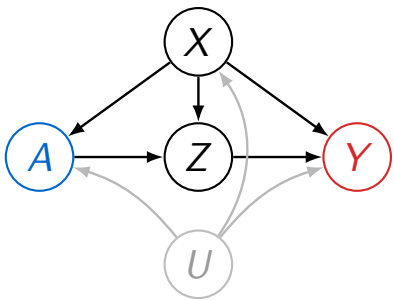
- x-axis: sample size n.
- y-axis: RMSE of $\hat{\tau}(X)$; lower is better.
- Fixed: $d = 30$ (Dimension of X).
- Nuisance noise: $n^{-1/4}$-scale perturbation added to nuisance fits.
- Curves: FD-PI, FD-DR, and full-noise FD-R.
- Readout: FD-R stays lowest; FD-DR improves with n but remains above FD-R.

Synthetic Results: Nuisance-Noise Sweep



- x-axis: nuisance noise level ρ.
- y-axis: RMSE of $\hat{\tau}(X)$; lower is better.
- Fixed setting: $n = 2000$ and $d = 30$.
- Larger ρ means deliberately noisier nuisance fits.
- Main readout: FD-PI and FD-DR degrade as ρ grows; FD-R remains comparatively stable.

FARS Setup



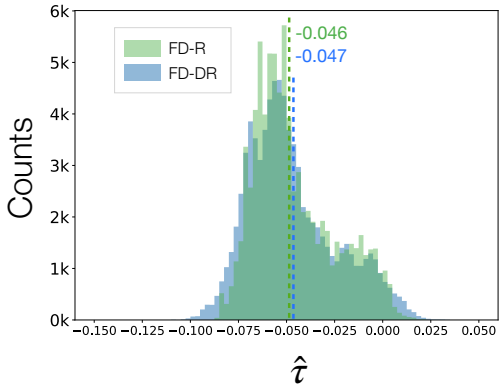
State seat-belt laws and occupant fatalities:

- **A** primary seat-belt law
- **Z** observed belt use
- **Y** occupant fatality
- **X** driver, occupant, time, vehicle, state-year
- **U** latent adoption and safety-culture factors

Front-door logic: law changes belt use, belt use changes fatality risk, and rich covariates help adjust the two observable pathway links.

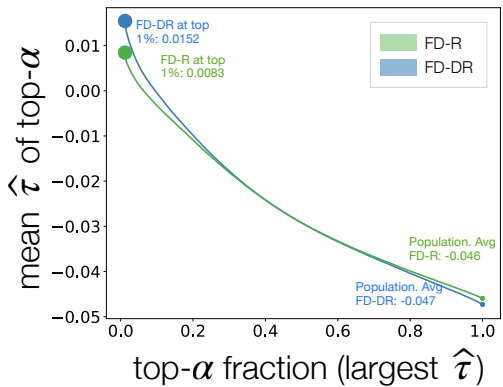
Data source: National Highway Traffic Safety Administration (2000), Fatality Analysis Reporting System (FARS).

FARS Results: Effect Distribution



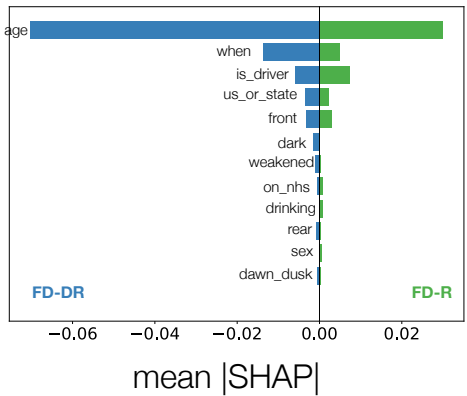
- Histogram of estimated individual effects $\hat{\tau}$.
- Green: FD-R. Blue: FD-DR.
- Both methods give similar effect distributions.
- Both averages are negative: about -0.046 to -0.047.
- Interpretation: the average-effect story is stable across learners.

FARS Results: Targeting Curve



- Sort observations by largest estimated $\hat{\tau}$.
- x-axis: top-α fraction under that ranking.
- y-axis: mean $\hat{\tau}$ within the selected fraction.
- Both learners show similar tail behavior.
- Interpretation: heterogeneity pattern is stable across FD-R and FD-DR.

FARS Results: Heterogeneity Drivers



- Bars show mean absolute SHAP contribution to $\hat{\tau}$.
- FD-DR and FD-R identify similar drivers.
- Age is dominant in both learners; driver role, road type, seating position, and lighting also contribute.
- Interpretation: method choice does not change the main empirical story.

References

Front-Door Identification And Estimation

- Jung, Y. *Debiased Front-Door Learners for Heterogeneous Effects*. arXiv:2509.22531.
- Pearl, J. (1995). *Causal diagrams for empirical research. Biometrika*, 82(4), 669–710.
- Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. (2020). *Robust inference on population indirect causal effects. JRSSB*, 82(1), 199–214.
- Guo, A., Benkeser, D., and Nabi, R. (2023). *Targeted machine learning for average causal effect estimation using the Front-Door functional*. arXiv:2312.10234.
- Wen, L., Sarvet, A., and Stensrud, M. (2024). *Causal effects of intervening variables in settings with unmeasured confounding. JMLR*, 25(345), 1–54.
- Jung, Y., Tian, J., and Bareinboim, E. (2024). *Unified Covariate Adjustment for Causal Inference*. NeurIPS.

Conditional FD And HTE Learners

- Chen, W., Chang, T., and Wiens, J. (2025). *Conditional Front-door Adjustment for Heterogeneous Treatment Assignment Effect Estimation Under Non-compliance*. CHIL/PMLR.
- Xu, L. and Gretton, A. (2022). *A Neural Mean Embedding Approach for Back-door and Front-door Adjustment*. ICLR.
- Xu, Z., Cheng, D., Li, J., Liu, J., Liu, L., and Yu, K. (2024). *Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder*. ICLR.
- Kennedy, E. H. (2023). *Towards optimal doubly robust estimation of heterogeneous causal effects. EJS*, 17(2), 3008–3049.
- Nie, X. and Wager, S. (2021). *Quasi-oracle estimation of heterogeneous treatment effects. Biometrika*, 108(2), 299–319.
- Foster, D. J. and Syrgkanis, V. (2023). *Orthogonal statistical learning. Annals of Statistics*, 51(3), 879–908.
- Chernozhukov, V. et al. (2018). *Double/debiased machine learning for treatment and structural parameters. Econometrics Journal*, 21(1), C1–C68.

Applied Example, Data, And Explanation

- Piccininni, M., Kurth, T., Audebert, H. J., and Rohmann, J. L. (2023). *The Effect of Mobile Stroke Unit Care on Functional Outcomes: An Application of the Front-door Formula. Epidemiology*, 34(5), 712–720.
- *Epidemiology*. (2024). *2024 Rothman Epidemiology Prize*. 35(4), 431.
- National Highway Traffic Safety Administration. (2000). *Fatality Analysis Reporting System (FARS) 2000 Data Files*. U.S. Department of Transportation.
- Lundberg, S. M. and Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. NeurIPS.

Thank you

Paper

arXiv:2509.22531

Code

GitHub