Examining Racial Disparities in Healthcare Expenditures via Causal Mediation Analysis

Xiaxian Ou and Razieh Nabi

Department of Biostatistics and Bioinformatics, Emory University

EMORY

Motivation

- **Racial disparities in healthcare expenditures** have been widely documented, yet the underlying factors remain complex and require further exploration.
- A multitude of interrelated factors complicates analysis: **socioeconomic status (SES), access to insurance, health behaviors, health status**.
- A flexible, nonparametric framework based on **counterfactual formalization** in path-specific analysis to identify and quantify sources of disparities is needed.
- Estimator derived from **efficient influence function** (EIF) and modeling technique involving **SuperLearner** are crucial for robust and reliable estimation.

Causal Path-Specific Effect Analysis

Estimand

A nested potential outcome:

$$\phi(r_0,r_1,r_2,r_3,r_4) := Y\Big(r_0, M_1(r_1), M_2\big(r_2, M_1(r_1)\big), M_3\big(r_3, M_1(r_1), M_2(r_2, M_1(r_1))\big),$$
$$M_4\Big(r_4, M_1(r_1), M_2\big(r_2, M_1(r_1)\big), M_3\big(r_3, M_1(r_1), M_2(r_2, M_1(r_1))\big)\Big)\Big),$$

The healthcare expenditures of individuals if they belonged to racial group $R = r_0$, with their SES (M_1), insurance (M_2), health behaviors (M_3), and health status (M_4) set to the natural levels they would have attained if they hypothetically belonged to racial groups r_1, r_2, r_3, and r_4, respectively, where $(r_0,r_1,r_2,r_3,r_4) \in \{0,1\}^4$.

Natural path-specific effects (PSEs):

$$\begin{aligned}
\rho_{R\to Y} &:= \mathbb{E}[\phi(1,0,0,0,0) - \phi(0,0,0,0,0)] ,\\
\rho_{R\to M_1 \rightsquigarrow Y} &:= \mathbb{E}[\phi(0,1,0,0,0) - \phi(0,0,0,0,0)] ,\\
\rho_{R\to M_2 \rightsquigarrow Y} &:= \mathbb{E}[\phi(0,0,1,0,0) - \phi(0,0,0,0,0)] ,\\
\rho_{R\to M_3 \rightsquigarrow Y} &:= \mathbb{E}[\phi(0,0,0,1,0) - \phi(0,0,0,0,0)] ,\\
\rho_{R\to M_4 \rightsquigarrow Y} &:= \mathbb{E}[\phi(0,0,0,0,1) - \phi(0,0,0,0,0)] .
\end{aligned}$$

Identification

Assumptions: (a) Consistency; (b) Conditional ignorability; (c) Positivity

Identification formula:

$$\rho_{R\to M_k \rightsquigarrow Y} = \int y \Big\{ dP(y \mid \overline{m}_4, R=0, x) \prod_{k=1}^{K} dP(m_k \mid \overline{m}_{k-1}, r_k, x) - dP(y \mid R=0, x) \Big\} dP(x) ,$$

$$\rho_{R\to Y} = \int y \Big\{ dP(y \mid \overline{m}_4, R=1, x) \prod_{k=1}^{K} dP(m_k \mid \overline{m}_{k-1}, R=0, x) - dP(y \mid R=0, x) \Big\} dP(x) .$$

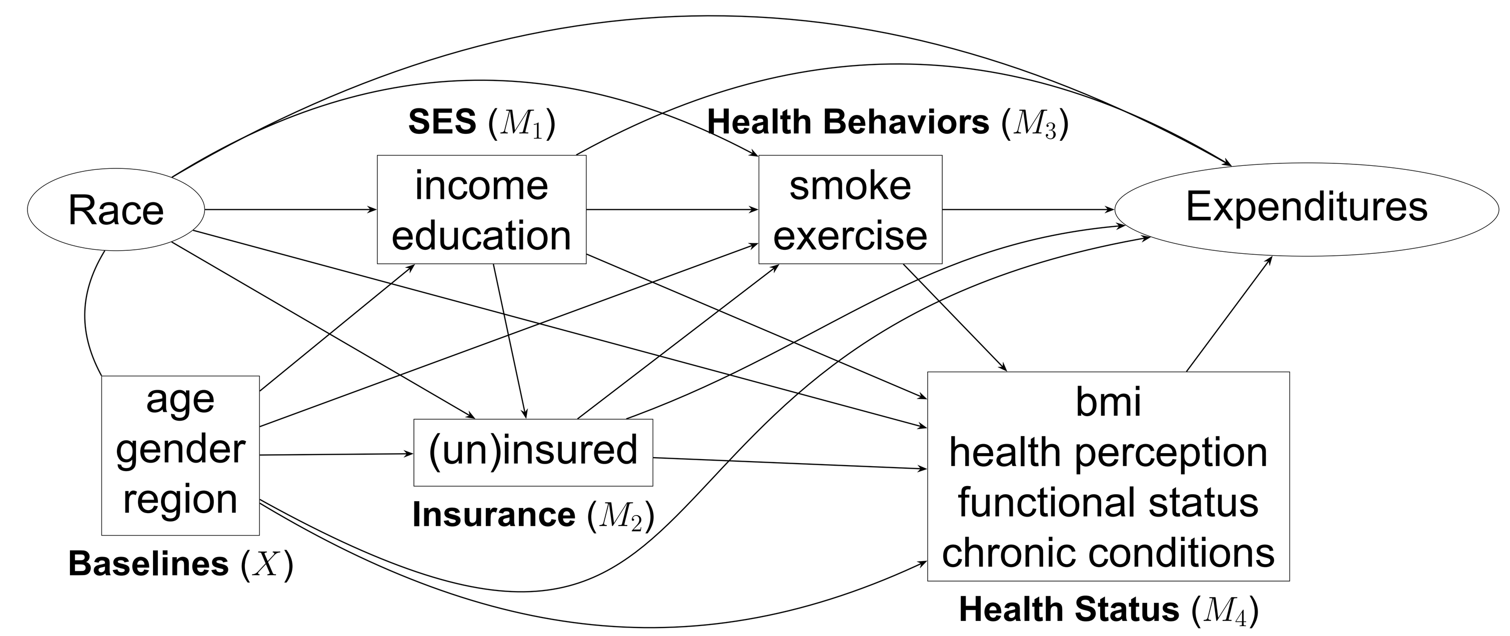
Multiply robust estimators

$$\begin{aligned}
\psi_{\rho_{R\to M_k \rightsquigarrow Y},n} =& \frac{1}{n}\sum_{i=1}^{n} \Bigg[\frac{(1-R_i)}{\hat{g}_0(0|x_i)} \frac{\hat{g}_k(1|\overline{m}_{k,i},x_i)}{\hat{g}_k(0|\overline{m}_{k,i},x_i)} \frac{\hat{g}_{k-1}(0|\overline{m}_{k-1,i},x_i)}{\hat{g}_{k-1}(1|\overline{m}_{k-1,i},x_i)} \big(Y_i - \hat{\mu}_k(\overline{m}_{k,i},0,x_i)\big) \\
&+ \frac{R_i}{\hat{g}_0(0|x_i)} \frac{\hat{g}_{k-1}(0|\overline{m}_{k-1,i},x_i)}{\hat{g}_{k-1}(1|\overline{m}_{k-1,i},x_i)} \big(\hat{\mu}_k(\overline{m}_{k,i},0,x_i) - \hat{\mathcal{B}}_k(\overline{m}_{k-1,i},1,x_i)\big) \\
&+ \frac{(1-R_i)}{\hat{g}_0(0|x_i)} \big(\hat{\mathcal{B}}_k(\overline{m}_{k-1,i},1,x_i) - \hat{\mathcal{C}}(\hat{\mathcal{B}}_k,r_1,x_i)\big) + \hat{\mathcal{C}}(\hat{\mathcal{B}}_k,r_1,x_i)\big) \Bigg]
\end{aligned}$$

$$\begin{aligned}
\psi_{\rho_{R\to Y},n} =& \frac{1}{n}\sum_{i=1}^{n} \Bigg[\frac{R_i}{\hat{g}_0(0|x_i)} \frac{\hat{g}_k(0|\overline{m}_{k,i},x_i)}{\hat{g}_k(1|\overline{m}_{k,i},x_i)} \big(Y_i - \hat{\mu}_k(\overline{m}_{k,i},1,x_i)\big) \\
&+ \frac{(1-R_i)}{\hat{g}_0(0|x_i)} \big(\hat{\mu}_k(\overline{m}_{k_i},1,x_i) - \hat{\mathcal{C}}(\hat{\mu}_{k_i},0,x_i)\big) + \hat{\mathcal{C}}(\hat{\mu}_{k,i},0,x_i) \Bigg]
\end{aligned}$$

- $\psi_{\rho_{R\to M_k \rightsquigarrow Y}} = E(\phi(r_0,r_1,r_2,r_3,r_4))$, where $r_k = 1, r_j = 0, j \neq k$
- $\mu_k(\overline{m}_k, r_0, x) = \mathbb{E}(Y|\overline{m}_k, r_0, x)$, $\mathcal{B}_k(\overline{m}_{k-1}, r_k, x) = \mathbb{E}(\mu_k(\overline{m}_k, r_0, x)|\overline{m}_{k-1}, r_k, x)$, $\mathcal{C}(\cdot, r_1, x) = \mathbb{E}(\cdot|r_1, x)$
- $g_k(r|\overline{m}_k, x) = P(r|\overline{m}_k, x)$

Graphical Representation



Empirical Analysis of MEPS Data

Medical Expenditures Panel Survey (MEPS) 2009: non-Hispanic Whites (9,830), non-Hispanic Blacks (3,905), Asians (1,431) and Hispanics (5,150)

Data challenge

1. **Zero-inflated right-skewed data**

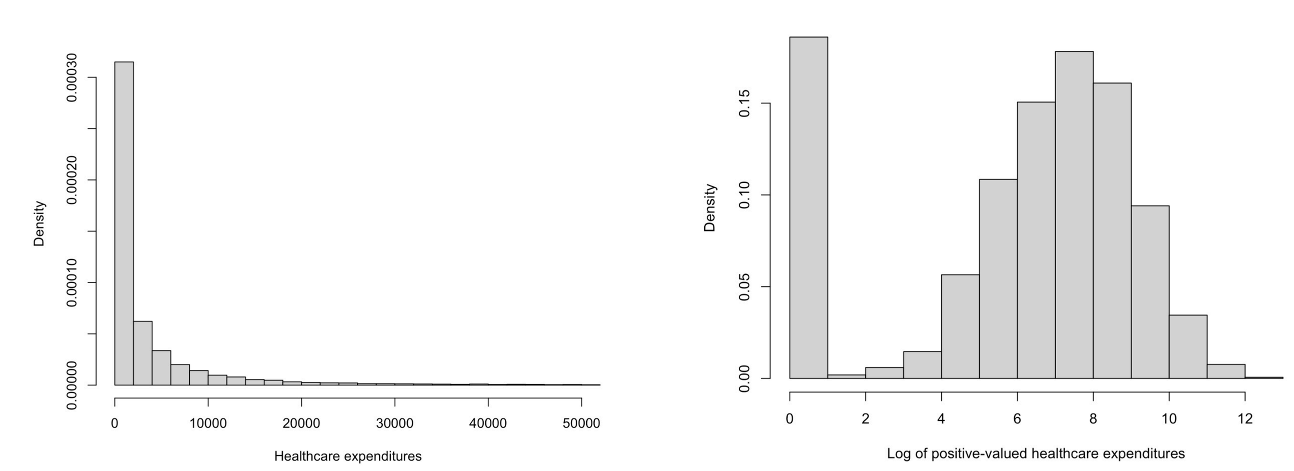


Figure 1. The original healthcare expenditures and the log transformation in positive data.

2. **Complex relation between treatment, mediators, and outcome**
   Naive use of ML may lead to large first-order bias of the plug-in estimator.

Approach

! Two-part model

- Part 1: The probability of a non-zero response, $P(Y > 0 \mid X)$,
- Part 2: The probability distribution of the positive responses $\mathbb{E}[\log(Y) \mid Y > 0, X]$

The conditional mean: $\mathbb{E}[\log(Y) \mid X] = P(Y > 0 \mid X) \times \mathbb{E}[\log(Y) \mid Y > 0, X]$.

! Transformation

- The estimated geometric mean G_n of ratio scale (e.g. total effect):

$$\psi_n(1) - \psi_n(0) = \frac{1}{n}\sum_{i=1}^{n}(\log(\hat{Y}_i(1)) - \log(\hat{Y}_i(0))) = \log\Big(\sqrt[n]{\frac{\hat{Y}_1(1)}{\hat{Y}_1(0)} \cdots \frac{\hat{Y}_n(1)}{\hat{Y}_n(0)}}\Big)$$

$$exp(\psi_n(1) - \psi_n(0)) = \sqrt[n]{\frac{\hat{Y}_1(1)}{\hat{Y}_1(0)} \cdots \frac{\hat{Y}_n(1)}{\hat{Y}_n(0)}} = G_n\left(\frac{\hat{Y}(1)}{\hat{Y}(0)}\right)$$

- Delta method:

$$\psi_n(1) - \psi_n(0) \sim N(\psi_0(1) - \psi_0(0), \frac{\sigma^2_{1,0}}{n})$$

$$exp(\psi_n(1) - \psi_n(0)) \sim N(exp(\psi_0(1) - \psi_0(0)), exp(\psi_0(1) - \psi_0(0))^2\frac{\sigma^2_{1,0}}{n})$$

! SuperLearner

- Binomial family: *glm*, *glm.interaction*, *randomForest*, *xgboost*, and *dbarts*
- Gaussian family: *glmnet*, *polymars*, *lm*, and *dbarts*

Results

Path	Effect(95%CI)	p value
Whites vs Blacks*		
$R \to M_1 \rightsquigarrow Y$	1.098(1.035~1.161)	0.001
$R \to M_2 \rightsquigarrow Y$	1.009(0.974~1.044)	0.606
$R \to M_3 \rightsquigarrow Y$	0.974(0.951~0.998)	0.035
$R \to M_4 \to Y$	1.035(0.963~1.107)	0.337
$R \to Y$	1.787(1.629~1.945)	0.000
Total effect	2.106(1.865~2.347)	0.000
Whites vs Asians*		
$R \to M_1 \rightsquigarrow Y$	0.945(0.834~1.055)	0.339
$R \to M_2 \rightsquigarrow Y$	1.054(0.992~1.116)	0.081
$R \to M_3 \rightsquigarrow Y$	0.982(0.927~1.036)	0.509
$R \to M_4 \to Y$	1.358(1.171~1.545)	0.000
$R \to Y$	2.521(2.159~2.883)	0.000
Total effect	2.805(2.304~3.306)	0.000
Whites vs Hispanics*		
$R \to M_1 \rightsquigarrow Y$	1.572(1.438~1.705)	0.000
$R \to M_2 \rightsquigarrow Y$	1.377(1.305~1.450)	0.000
$R \to M_3 \rightsquigarrow Y$	1.089(1.020~1.159)	0.009
$R \to M_4 \to Y$	1.436(1.307~1.564)	0.000
$R \to Y$	2.044(1.865~2.223)	0.000
Total effect	4.647(4.143~5.150)	0.000

Path	Effect(95%CI)	p value
Blacks vs Asians*		
$R \to M_1 \rightsquigarrow Y$	0.837(0.692~0.981)	0.043
$R \to M_2 \rightsquigarrow Y$	1.024(0.947~1.101)	0.531
$R \to M_3 \rightsquigarrow Y$	0.970(0.917~1.023)	0.271
$R \to M_4 \to Y$	1.475(1.243~1.708)	0.000
$R \to Y$	1.111(0.917~1.305)	0.237
Total effect	1.297(1.008~1.585)	0.022
Blacks vs Hispanics*		
$R \to M_1 \rightsquigarrow Y$	1.268(1.194~1.343)	0.000
$R \to M_2 \rightsquigarrow Y$	1.486(1.384~1.588)	0.000
$R \to M_3 \rightsquigarrow Y$	1.053(1.006~1.099)	0.022
$R \to M_4 \to Y$	1.367(1.183~1.552)	0.000
$R \to Y$	0.988(0.910~1.066)	0.770
Total effect	2.111(1.791~2.431)	0.000
Asians vs Hispanics*		
$R \to M_1 \rightsquigarrow Y$	1.960(1.706~2.213)	0.000
$R \to M_2 \rightsquigarrow Y$	1.342(1.221~1.463)	0.000
$R \to M_3 \rightsquigarrow Y$	0.998(0.978~1.018)	0.846
$R \to M_4 \to Y$	0.811(0.716~0.906)	0.000
$R \to Y$	0.988(0.912~1.064)	0.760
Total effect	1.884(1.541~2.227)	0.000

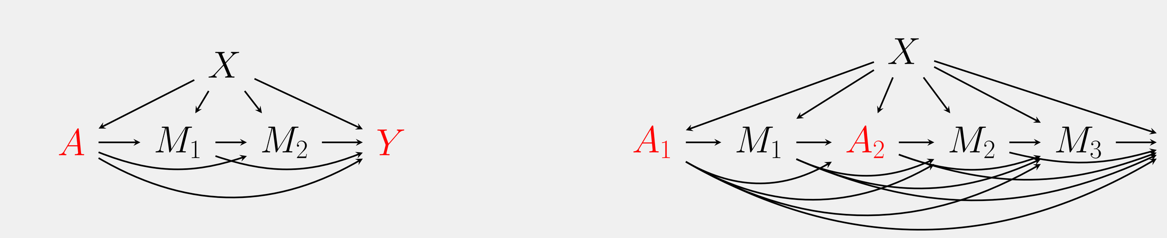
★ Total effects were significant in all race comparisons

★ The effects via SES and health status were significant in five comparisons.

★ The direct effects were significant in the comparisons between Whites and any minority.

R Package: *flexPaths*

1. Flexible model size: Flexible number of treatments and mediators.
2. Flexible decomposition: The Natural PSEs and the Cumulative PSEs.
3. Flexible pathways: The PSE through flexible identified pathway(s).
4. Flexible models: *glm/lm, dbarts, SuperLearner* and user-extended model.
5. Flexible estimators: Inverse Probability Weighting (IPW), plug-in G-computation, and EIF estimator.



(a) Single treatment (b) Multiple treatments

- A nested potential outcome

$$\phi(r_{11}, r_{12}, r_{10}) \qquad \phi\begin{pmatrix} r_{11}, & r_{12}, & r_{13}, & r_{10}, \\ & r_{22}, & r_{23}, & r_{20} \end{pmatrix}$$

$r_{ij} \in \{0,1\}$: the counterfactual value of i_{th} treatment for j_{th} mediator.
e.g. direct effect:

$$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 0 & 1 \\ na & 0 & 0 & 1 \end{bmatrix}$$

References

[1] J. Pearl, "Direct and Indirect Effects," in *The Seventeenth Conference*, (San Francisco, CA: Morgan Kaufmann), pp. 411–420, 2001.

[2] X. Zhou, "Semiparametric Estimation for Causal Mediation Analysis with Multiple Causally Ordered Mediators," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, pp. 794–821, July 2022.