On Counterfactual Metrics for Social Welfare: Incentives, Ranking, and Information Asymmetry

Serena Wang
Joint with Stephen Bates, P. M. Aronow, Michael I. Jordan
CIRM, December 14, 2023

Berkeley
UNIVERSITY OF CALIFORNIA
Electrical Engineering
and Computer Sciences

Charity Navigator

List of Best Highly Rated Charities

ECHO of Brandon
Brandon , FL

Large-sized | Basic and emergency aid

100%

Donate

Feeding America Eastern Wisconsin
Milwaukee , WI

Super-sized | Free goods distribution

100%

10 Top Real Estate Agents in Berkeley by Sales Transactions

AGENT NAME	BROKERAGE	SALES IN PRIOR 12 MONTHS*	PRICE OF SALES OVER 24 MONTHS*
Brett J	Real Estate Experts	562	$1,160,907.75 - $1,934,846.25
Anthony C	Golden Gate Sotheby's	59	$1,235,730.00 - $2,059,550.00
Andrea G	COMPASS	54	$1,026,346.50 - $1,710,577.50

University of California Berkeley

United States | Berkeley

#19 in **Best Universities for Artificial Intelligence**
#4 in **Best Global Universities**

The University of California—Berkeley is situated roughly 15 miles from San Francisco in what is known as the Bay Area... Read More »

Subject Score
82.3

Global Score
88.7

Enrollment
40,921

Charity Navigator

List of Best Highly Rated Charities

Wisconsin
Milwaukee , WI

Super-sized | Free goods distribution

100%

Best Hospitals Honor Roll

Each year, U.S. News ranks hospitals in 15 specialties and 20 procedures and conditions. Here are the top 20 highest rated hospitals in the U.S.

2022-2023 Honor Roll Rankings

- #1 Mayo Clinic
  Rochester, MN
- #2 Cedars-Sinai Medical Center
  Los Angeles, CA
- #3 NYU Langone Hospitals
  New York, NY

SEE FULL RANKINGS LIST »

10 Top Real Estate Agents in Berkeley by Sales Transactions

AGENT NAME	BROKERAGE	SALES IN PRIOR 12 MONTHS*	PRICE OF SALES OVER 24 MONTHS*
J	Real Estate Experts	562	$1,160,907.75 - $1,934,846.25
ony C	Golden Gate Sotheby's	59	$1,235,730.00 - $2,059,550.00

Member Login

Cal Hospital Compare

Learning Center | About | My Hospitals

Ratings You Can Trust.

Cal Hospital Compare makes it easy to find and compare the quality of hospitals in California.

Compare Hospitals | Search by Zip Code, City, or Hospital Name

C-Section Honor Roll Hospitals Announced

One hundred and eleven California hospitals met or surpassed a federal target aimed at reducing Cesarean births (C-sections) for first-time mothers with low-risk pregnancies.

Browse featured content

Unive

United

#19 in **Best Universities for Artificial Intelligence**
#4 in **Best Global Universities**

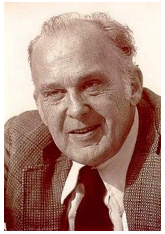
The University of California—Berkeley is situated roughly 15 miles from San Francisco in what is known as the Bay Area... Read More »

Global Score
88.7

Enrollment
40,921

Metrics often don’t match goals

Goodhart (1975) *(Economist)*: “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.”



Campbell (1979) *(Social Psychologist)*: “The more any quantitative social indicator is used for social decision-making, [...] the more apt it will be to distort and corrupt the social processes it is intended to monitor.”

Strathern (1997) *(Anthropologist)*: “When a measure becomes a target, it ceases to be a good measure.”

Domain

Monetary policy and finance

Social programs

Political processes

Education

Business and management

Behind every algorithm is a metric.

1990: New York

Hospital

Mortality rate

%

NEW YORK STATE | Department of Health

Health Department

1990: New York

Hospital

Mortality rate

%

NEW YORK STATE | Department of Health

Health Department

"**More severely ill patients experienced dramatically worsened health outcomes**" (Dranove et al. 2003).

Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is more information better? The effects of "report cards" on health care providers. *Journal of Political Economy*, 111(3), 555-588.

1990: New York

Patients

Selection

✓ ✗

Hospital

Mortality rate

%

NEW YORK STATE | Department of Health

Health Department

"**More severely ill patients experienced dramatically worsened health outcomes**" (Dranove et al. 2003).

Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is more information better? The effects of "report cards" on health care providers. *Journal of Political Economy*, 111(3), 555-588.

Patients

Selection

✓ ✗

Hospital

Mortality rate

%

Centers for Medicare and Medicaid Services (CMS)

Best Hospitals Honor Roll

Each year, U.S. News ranks hospitals in 15 specialties and 20 procedures and conditions. Here are the top 20 highest rated hospitals in the U.S.

2022-2023 Honor Roll Rankings

- #1 Mayo Clinic, Rochester, MN
- #2 Cedars-Sinai Medical Center, Los Angeles, CA
- #3 NYU Langone Hospitals, New York, NY

SEE FULL RANKINGS LIST »

Students

Selection

✓ ✗

University

Student success rates (e.g. graduation, employment)

%

BEST COLLEGES
U.S.News & WORLD REPORT
RANKINGS

US News and World Report; Accrediting agencies

University of California Berkeley

United States | Berkeley

#19 in **Best Universities for Artificial Intelligence**
#4 in **Best Global Universities**

The University of California—Berkeley is situated roughly 15 miles from San Francisco in what is known as the Bay Area... Read More »

Subject Score
82.3

Global Score
88.7

Enrollment
40,921

Home buyers/sellers

Selection

✓ ✗

Real estate agents

Sales volume

%

Zillow

Zillow, Yelp, Google,
US News and World, etc.

10 Top Real Estate Agents in Berkeley by Sales Transactions

AGENT NAME	BROKERAGE	SALES IN PRIOR 12 MONTHS*	PRICE OF SALES OVER 24 MONTHS*
Brett J	Real Estate Experts	562	$1,160,907.75 - $1,934,846.25
Anthony C	Golden Gate Sotheby's	59	$1,235,730.00 - $2,059,550.00
Andrea G	COMPASS	54	$1,026,346.50 - $1,710,577.50

Key Questions

Question 1 (Incentives): How do we design quality metrics that lead to better treatment incentives?

Key Questions

Question 1 (Incentives): How do we design quality metrics that lead to better treatment incentives?

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Key Questions

Question 1 (Incentives): How do we design quality metrics that lead to better treatment incentives?

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

Related work

"Cream skimming" [Culyer & Newhouse, 2000; Koning & Heinrich, 2013, Glazer & McGuire, 2000]

Empirical; theoretical

Counterfactual metrics for better incentives in modern ranking systems driven by data/ML.

Culyer, A. J., & Newhouse, J. P. (Eds.). (2000). *Handbook of health economics*. Elsevier.
Koning, P., & Heinrich, C. J. (2013). Cream-skimming, parking and other intended and unintended effects of high-powered, performance-based contracts. *Journal of Policy Analysis and Management,* 32(3), 461-483.
Glazer, J. & McGuire, T. G. (2000). "Optimal risk adjustment in markets with adverse selection: an application to managed care." *American Economic Review* 90(4): 1055-71.

Related work

"Cream skimming" [Culyer & Newhouse, 2000; Koning & Heinrich, 2013, Glazer & McGuire, 2000]

Policy learning [Manski, 2008; Kitagawa and Tetenov, 2018; Athey and Wager, 2021]

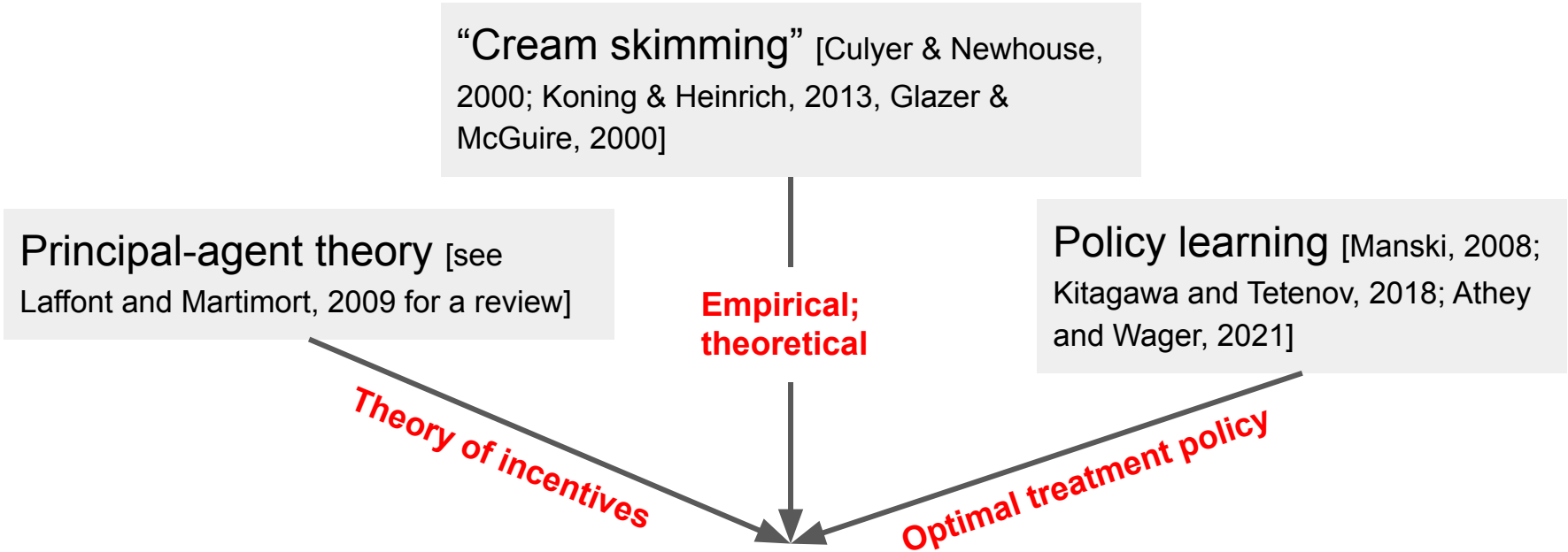
Empirical; theoretical

Optimal treatment policy

Counterfactual metrics for better incentives in modern ranking systems driven by data/ML.

Culyer, A. J., & Newhouse, J. P. (Eds.). (2000). *Handbook of health economics*. Elsevier.
Koning, P., & Heinrich, C. J. (2013). Cream-skimming, parking and other intended and unintended effects of high-powered, performance-based contracts. *Journal of Policy Analysis and Management,* 32(3), 461-483.
Glazer, J. & McGuire, T. G. (2000). "Optimal risk adjustment in markets with adverse selection: an application to managed care." *American Economic Review* 90(4): 1055-71.
Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.
Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133-161.
Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2), 591-616.

Related work



Counterfactual metrics for better incentives in modern ranking systems driven by data/ML.

Culyer, A. J., & Newhouse, J. P. (Eds.). (2000). *Handbook of health economics*. Elsevier.
Koning, P., & Heinrich, C. J. (2013). Cream-skimming, parking and other intended and unintended effects of high-powered, performance-based contracts. *Journal of Policy Analysis and Management,* 32(3), 461-483.
Glazer, J. & McGuire, T. G. (2000). "Optimal risk adjustment in markets with adverse selection: an application to managed care." *American Economic Review* 90(4): 1055-71.
Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.
Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133-161.
Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2), 591-616.
Laffont, J. J., & Martimort, D. (2009). *The Theory of Incentives*. Princeton University Press.

Notation: potential outcomes (Neyman–Rubin)

Patient covariates: $X_i \in \mathbb{R}^d$

Treatment assignment: $T_i \in \{0, 1\}$

Potential outcomes if untreated or treated: $Y_i(0), Y_i(1)$

Observed outcome: $Y_i = Y_i(T_i)$

Notation: potential outcomes (Neyman–Rubin)

Patient covariates: $X_i \in \mathbb{R}^d$

Treatment assignment: $T_i \in \{0, 1\}$

Potential outcomes if untreated or treated: $Y_i(0), Y_i(1)$

Observed outcome: $Y_i = Y_i(T_i)$

Conditional average potential outcomes:

$$\mu_1(x) = E[Y_i(1)|X_i = x]$$
$$\mu_0(x) = E[Y_i(0)|X_i = x]$$

Principal-agent model

Principal (Health Dept)

Agent (Hospital)

Principal-agent model

Decentralized information: Agent knows more about task and effort than principal.

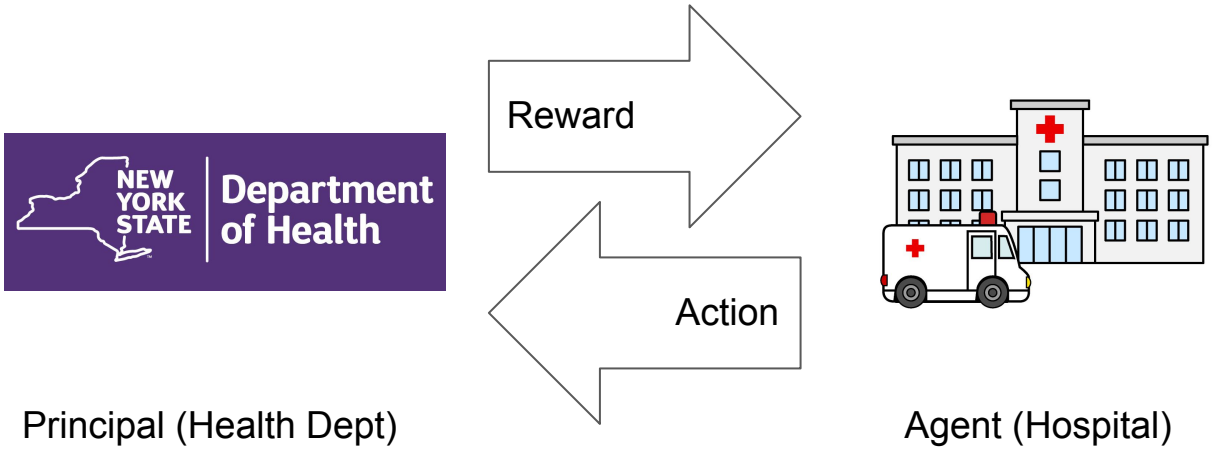
Principal (Health Dept)

Agent (Hospital)

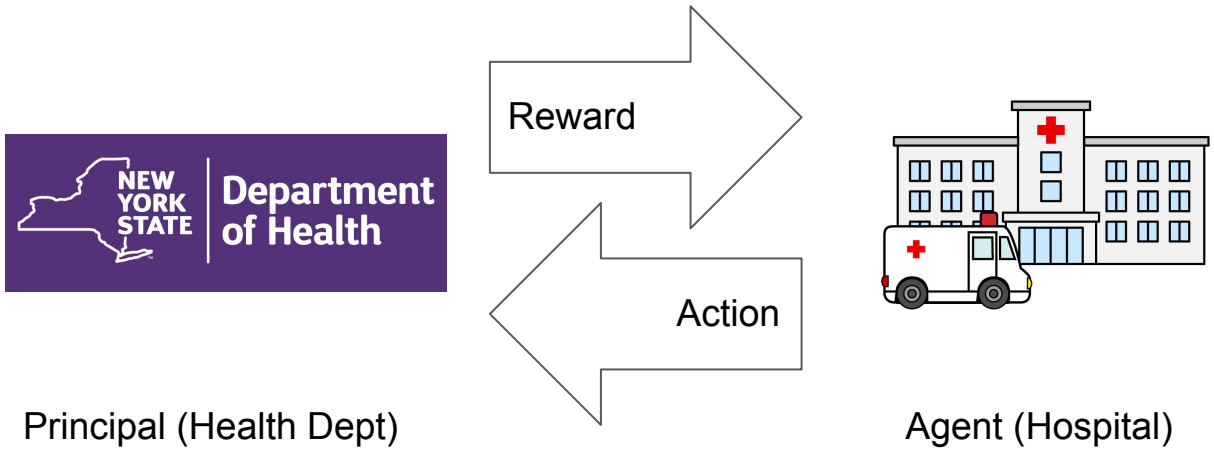
Principal-agent model

Decentralized information: Agent knows more about task and effort than principal.

Decentralized interests: Principal sets contract that determines agent's reward.



Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?



Principal-agent setup

Principal (Health Dept)

Agent (Hospital)

Principal-agent setup

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

(Possibly missing if $T_i = 0$)

Agent (Hospital)

Principal-agent setup

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

(Possibly missing if $T_i = 0$)

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Principal-agent setup

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

(Possibly missing if $T_i = 0$)

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Principal-agent setup

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

(Possibly missing if $T_i = 0$)

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Knows

Conditional average potential outcomes

$\mu_0(X_i), \mu_1(X_i)$

Principal-agent setup

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

(Possibly missing if $T_i = 0$)

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Knows

Conditional average potential outcomes

$\mu_0(X_i), \mu_1(X_i)$

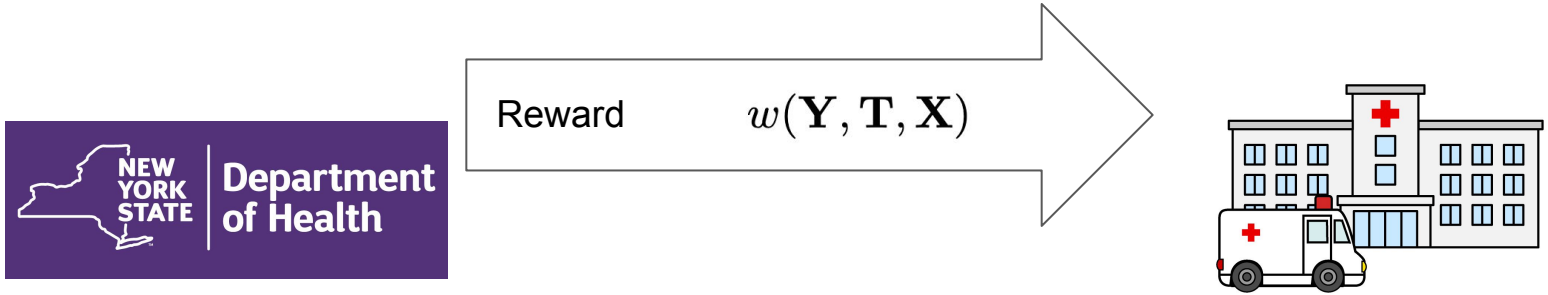
Chooses

Treatment rule

$\pi(x) = P(T_i = 1 | X_i = x)$

Game

1. Principal (Health Dept) chooses reward function w

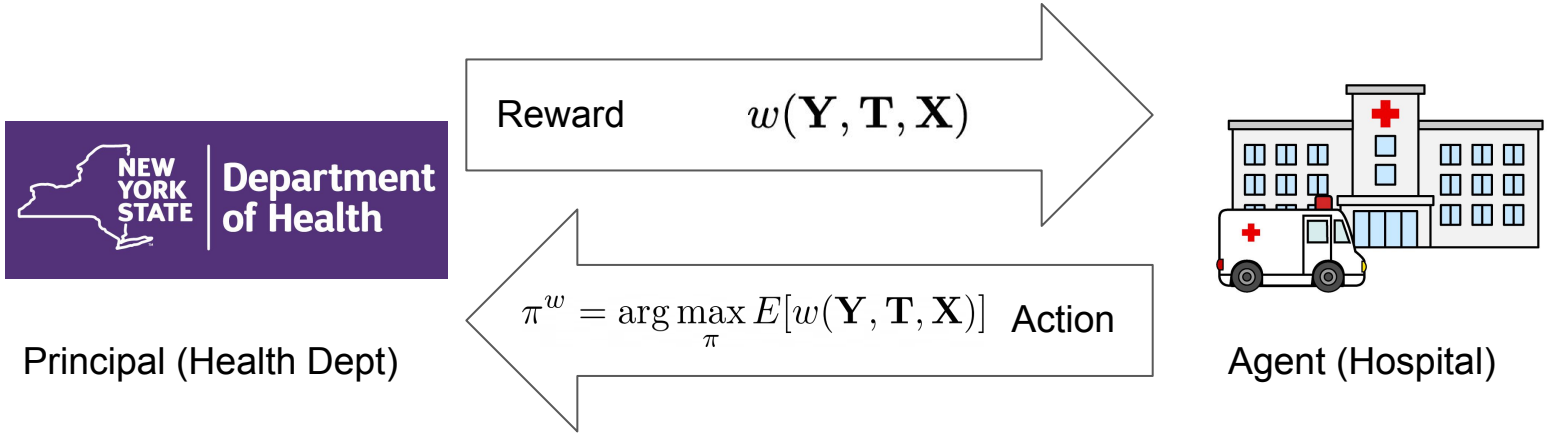


Principal (Health Dept)

Agent (Hospital)

Game

1. Principal (Health Dept) chooses reward function w
2. Agent (Hospital) best responds with treatment rule π^w



Defining welfare and regret

Principal's goal: design a reward to maximize **welfare effect** (Manski, 2009):

$$V(\pi) = E[\underbrace{Y_i(T_i^\pi)}_{\text{Outcomes under treatment rule}} - \underbrace{Y_i(0)}_{\text{Outcomes if no one gets treated}}]$$

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

Defining welfare and regret

Principal's goal: design a reward to maximize **welfare effect** (Manski, 2009):

$$V(\pi) = E[\underbrace{Y_i(T_i^\pi)}_{\text{Outcomes under treatment rule}} - \underbrace{Y_i(0)}_{\text{Outcomes if no one gets treated}}]$$

Regret:

$$R(\pi^w) = \underbrace{\max_\pi V(\pi)}_{\text{Welfare effect of best policy}} - \underbrace{V(\pi^w)}_{\text{Welfare effect of policy optimized under reward function } w}$$

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

Incentive problems with the status quo

Reward function 1: Status quo mortality rate, “Average Treated Outcome”

$$w_{\mathrm{ATO}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \frac{\sum_{i=1}^{n} Y_i T_i^{\pi}}{\sum_{i=1}^{n} T_i^{\pi}}$$

Incentive problems with the status quo

Reward function 1: Status quo mortality rate, “Average Treated Outcome”

$$w_{\mathrm{ATO}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \frac{\sum_{i=1}^{n} Y_i T_i^{\pi}}{\sum_{i=1}^{n} T_i^{\pi}}$$

Agent’s best response: **Treat only the patients with the highest expected treated outcome.**

$$\pi^{w_{\mathrm{ATO}}}(x) = \mathbb{1}(x \in \arg\max_x \mu_1(x) \text{ and } \mu_1(x) > 0)$$

Proposition 1: Regret is **unbounded!**

Change 1: reward benefit from treatment

Reward function 2: “Average Treatment Effect on the Treated” (ATT)

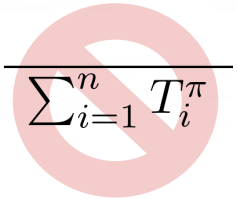
$$w_{\mathrm{ATT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}}{\sum_{i=1}^{n} T_i^{\pi}}$$

Estimated untreated counterfactual

Change 2: reward total effect

Reward function 3 (proposed): Aligned with welfare, “Total Treatment Effect”

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$



$$\frac{}{\sum_{i=1}^{n} T_i^{\pi}}$$

Change 2: No denominator (“total effect” instead of “average effect”)

Change 2: reward total effect

Reward function 3 (proposed): Aligned with welfare, “Total Treatment Effect”

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Agent’s best response: **Treat all patients with a positive treatment effect.**

$$\pi^{w_{\mathrm{TT}}}(x) = \mathbb{1}(\tau(x) > 0)$$

Proposition 3: Regret is **zero as long as** $\hat{\mu}_0(x)$ **is unbiased**

Change 2: reward total effect

Reward function 3 (proposed): Aligned with welfare, “Total Treatment Effect”

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Agent’s best response: **Treat all patients with a positive treatment effect.**

$$\pi^{w_{\mathrm{TT}}}(x) = \mathbb{1}(\tau(x) > 0)$$

Proposition 3: Regret is **zero as long as** $\hat{\mu}_0(x)$ **is unbiased** ← **Causal inference problem**

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Key result: "total treatment effect" reward function

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Reward based on agent treated data

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Key result: "total treatment effect" reward function

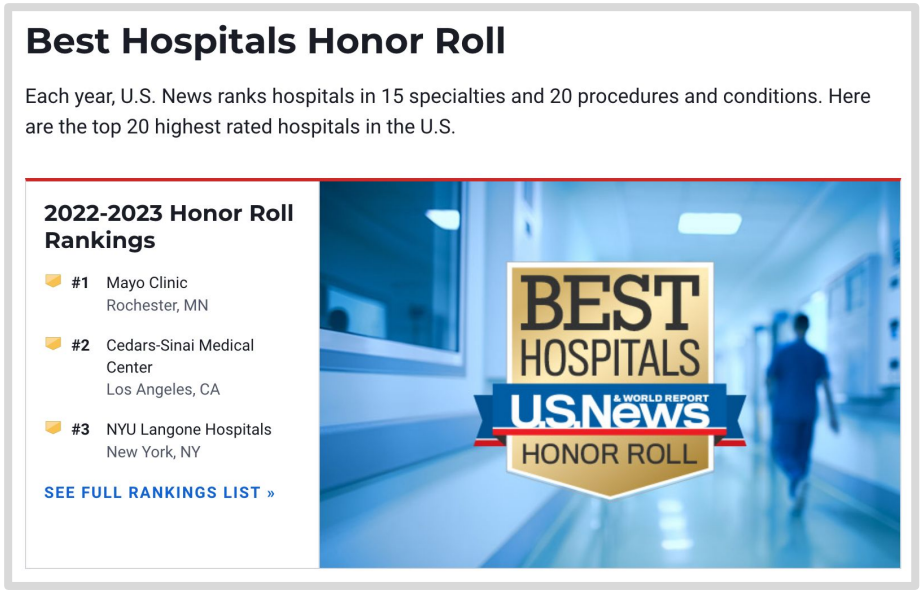
$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

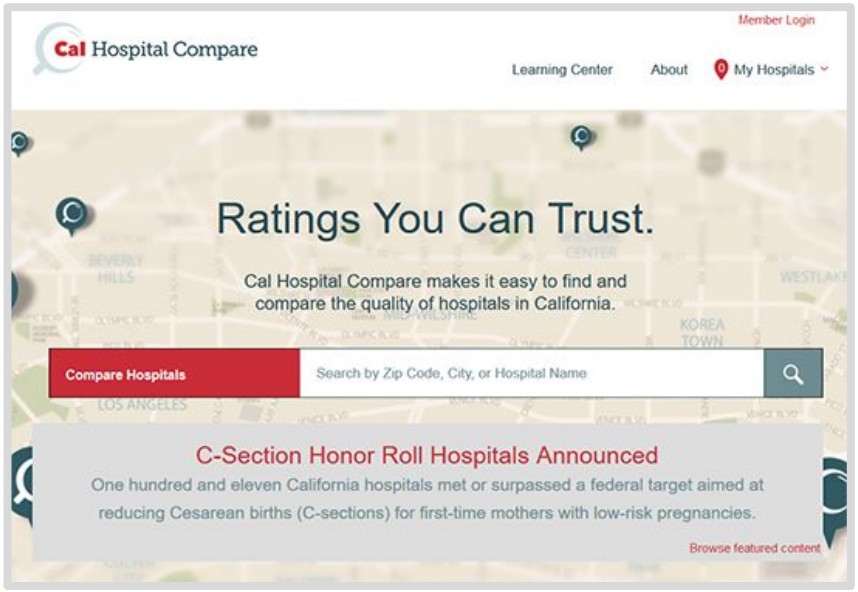
Learn from combo of historical data and agent untreated data

Reward based on agent treated data

Key questions revisited

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?





Do rankings behave reasonably?

Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.

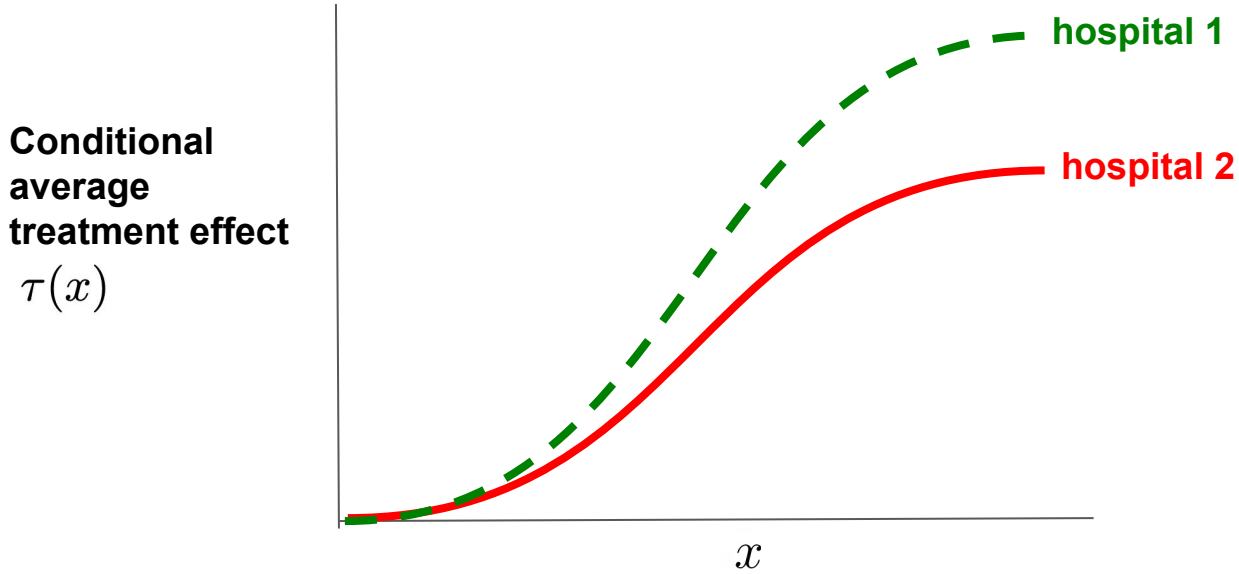
Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

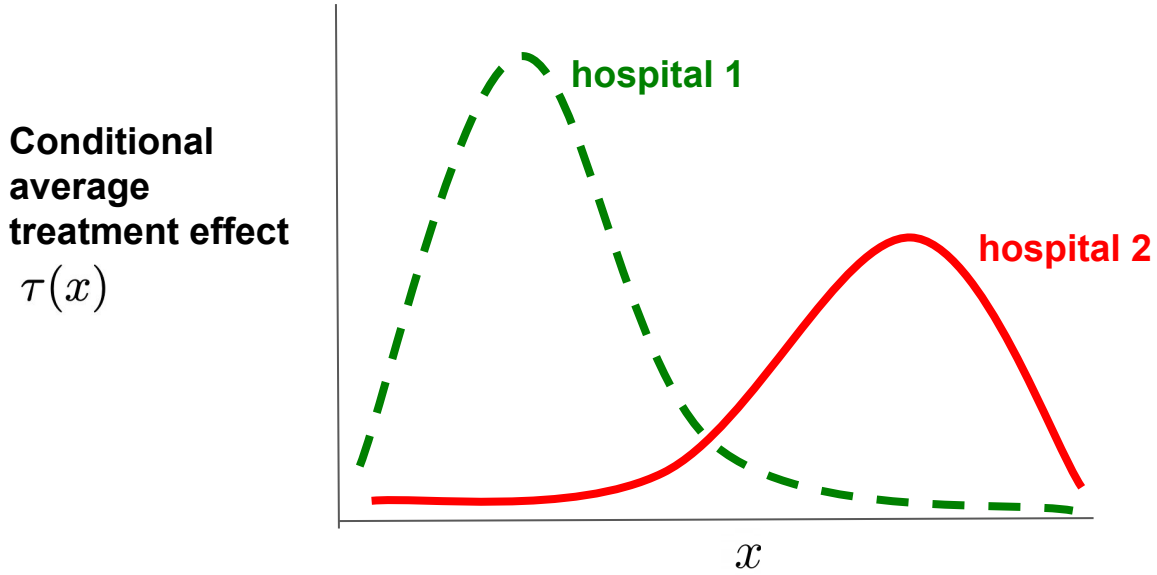
Better hospitals should be ranked higher (formal)

Definition 1 (Uniform Rank Preservation): If **hospital 1** is uniformly better than **hospital 2** for all x, then $w_1 > w_2$.



Better hospitals should be ranked higher (formal)

Definition 2 (Relative Rank Preservation): If **hospital 1** is better than **hospital 2** on average for a reference population, then $w_1 > w_2$.



Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

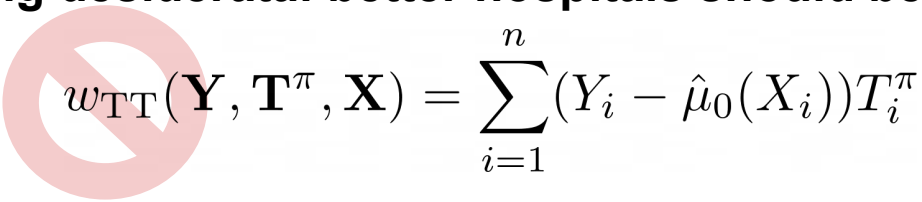
Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.



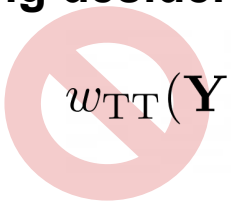
$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Problems with this reward function when comparing hospitals j and k:

- Advantages larger hospitals ($n_k > n_j$)

Do rankings behave reasonably?

Basic ranking desiderata: better hospitals should be ranked higher.



$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Problems with this reward function when comparing hospitals j and k:

- Advantages larger hospitals ($n_k > n_j$)
- Different hospitals serve diverse patient populations (Kim et al. 2022)
  Advantages "easier" population $X^{(k)}$ vs. $X^{(j)}$.

Kim, H., Mahmood, A., Hammarlund, N. E., & Chang, C. F. (2022). Hospital value-based payment programs and disparity in the United States: A review of current evidence and future perspectives. *Frontiers in Public Health*, 10, 882715.

Solution: reweight to reference population

$$w_k(\mathbf{Y}, \mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n_k} \left(Y_i - \hat{\mu}_0(X_i^{(k)}) \right) T_i^{\pi} \left(\underbrace{\frac{1}{n_k} \frac{p_0(X_i^{(k)})}{p_k(X_i^{(k)})}} \right)$$

Reweighting term

Solution: reweight to reference population

$$w_k(\mathbf{Y}, \mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n_k} \left(Y_i - \hat{\mu}_0(X_i^{(k)}) \right) T_i^{\pi} \left(\frac{1}{n_k} \frac{p_0(X_i^{(k)})}{p_k(X_i^{(k)})} \right)$$

Solution: reweight to reference population

density for reference population X_0

$$w_k(\mathbf{Y}, \mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n_k} \left(Y_i - \hat{\mu}_0(X_i^{(k)})\right) T_i^{\pi} \left(\frac{1}{n_k} \frac{p_0(X_i^{(k)})}{p_k(X_i^{(k)})}\right)$$

density for hospital population $X^{(k)}$

(See paper for Radon-Nikodym derivative form of the likelihood ratio.)

Solution: reweight to reference population

density for reference population X_0

$$w_k(\mathbf{Y}, \mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n_k} \left(Y_i - \hat{\mu}_0(X_i^{(k)})\right) T_i^{\pi} \left(\frac{1}{n_k} \frac{p_0(X_i^{(k)})}{p_k(X_i^{(k)})}\right)$$

density for hospital population $X^{(k)}$

Theorem 1 (Ranking Desiderata Satisfied): Hospitals will be ranked higher if they are (i) uniformly better at treating all patients, or (ii) better on average for reference population X_0.

Solution: reweight to reference population

density for reference population X_0

$$w_k(\mathbf{Y}, \mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n_k} \left(Y_i - \hat{\mu}_0(X_i^{(k)})\right) T_i^{\pi} \left(\frac{1}{n_k} \frac{p_0(X_i^{(k)})}{p_k(X_i^{(k)})}\right)$$

density for hospital population $X^{(k)}$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

General incentive-aligned form

Any positive function; policymaker's choice!

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} g(X_i)$$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

Accounting for policy desiderata

Any positive function; policymaker's choice!

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} \, g(X_i)$$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

Key result: this functional form **decouples** ranking desiderata from incentive alignment! A designer can optimize g for any ranking **policy desiderata**.

Accounting for policy desiderata

Any positive function; policymaker's choice!

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} \, g(X_i)$$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

Key result: this functional form **decouples** ranking desiderata from incentive alignment! A designer can optimize g for any ranking **policy desiderata**.

Example design choice: Selecting a reference population X_0:

General user population (default)

Accounting for policy desiderata

Any positive function; policymaker's choice!

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} \, g(X_i)$$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

Key result: this functional form **decouples** ranking desiderata from incentive alignment! A designer can optimize g for any ranking **policy desiderata**.

Example design choice: Selecting a reference population X_0:

General user population (default)

Tailor to age/gender/race/income/etc.

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Are rankings based on quality metrics useful for patients?

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Are rankings based on quality metrics useful for patients?

Key result: incentive-aligned function class

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} g(X_i)$$

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Are rankings based on quality metrics useful for patients?

Key result: incentive-aligned function class

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} g(X_i)$$

Reward + Rank with this

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Are rankings based on quality metrics useful for patients?

Key result: incentive-aligned function class

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} g(X_i)$$

Reward + Rank with this

Learn this

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Are rankings based on quality metrics useful for patients?

Key result: incentive-aligned function class

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} g(X_i)$$

Reward + Rank with this

Learn this

Choose this

Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

"Providers may be able to improve their ranking by selecting patients on the basis of characteristics that are **unobservable to the analysts but predictive of good outcomes**" (Dranove et al., 2003).

Setup under information asymmetry

Principal (Health Dept)

Agent (Hospital)

Setup under information asymmetry

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

Agent (Hospital)

Setup under information asymmetry

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Setup under information asymmetry

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

$\mathbf{U} = U_1, ..., U_n$

Setup under information asymmetry

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$
$\mathbf{U} = U_1, ..., U_n$

Knows

Expected potential outcomes

$E[Y_i(1)|X_i, U_i]$ $E[Y_i(0)|X_i, U_i]$

Setup under information asymmetry

Principal (Health Dept)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$

Treatments $\mathbf{T} = T_1, ..., T_n$

Outcomes $\mathbf{Y} = Y_1, ..., Y_n$

Chooses

Reward function $w(\mathbf{Y}, \mathbf{T}, \mathbf{X})$

Agent (Hospital)

Observes

Covariates $\mathbf{X} = X_1, ..., X_n$
$\mathbf{U} = U_1, ..., U_n$

Knows

Expected potential outcomes

$E[Y_i(1)|X_i, U_i]$ $E[Y_i(0)|X_i, U_i]$

Chooses

Treatment rule

$\pi(x, u) = P(T_i = 1|X_i = x, U_i = u)$

Consequences of information asymmetry

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Consequences of information asymmetry

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Confounding bias may or may not be present in auxiliary data:

$$E[\hat{\mu}_0(x)] \neq \mu_0(x)$$

Consequences of information asymmetry

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Confounding bias may or may not be present in auxiliary data:

$E[\hat{\mu}_0(x)] \neq \mu_0(x)$

Unmeasured heterogeneity:
Only measuring $\mu_0(x)$,
but agent chooses $\pi(x, u)$.

Consequences of information asymmetry

$$w_{\mathrm{TT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi}$$

Confounding bias may or may not be present in auxiliary data:

$E[\hat{\mu}_0(x)] \neq \mu_0(x)$

Unmeasured heterogeneity:
Only measuring $\mu_0(x)$,
but agent chooses $\pi(x, u)$.

Key result: When there is information asymmetry, unconfoundedness is not enough!
There is still regret, even if $\hat{\mu}_0(x)$ is unbiased.

Bounds on regret under information asymmetry

Bounds on regret under information asymmetry

Assumption (Bounded Unmeasured Heterogeneity): the effect of U on the untreated potential outcome given X is bounded:

$$E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$$

Bounds on regret under information asymmetry

Assumption (Bounded Unmeasured Heterogeneity): the effect of U on the untreated potential outcome given X is bounded:

$$E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$$

Upper bound: the regret under reward w_{TT} is upper bounded by the degree of unmeasured heterogeneity: $R(\pi^{w_{\text{TT}}}) \leq 2\gamma_{\text{marg}}$

Bounds on regret under information asymmetry

Assumption (Bounded Unmeasured Heterogeneity): the effect of U on the untreated potential outcome given X is bounded:

$$E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$$

Upper bound: the regret under reward w_{TT} is upper bounded by the degree of unmeasured heterogeneity: $R(\pi^{w_{\text{TT}}}) \leq 2\gamma_{\text{marg}}$

Lower bound: the regret is lower bounded by γ_{marg}.

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Key results: Reward total treatment effect by estimating the untreated counterfactual.

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Key results: Reward total treatment effect by estimating the untreated counterfactual.

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Key results: Not necessarily, but reweighting preserves incentive alignment.

Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

Key questions revisited

Question 1 (Incentives): How do we design quality metrics that lead to better hospital incentives?

Key results: Reward total treatment effect by estimating the untreated counterfactual.

Question 2 (Ranking): Do rankings based on those quality metrics behave reasonably?

Key results: Not necessarily, but reweighting preserves incentive alignment.

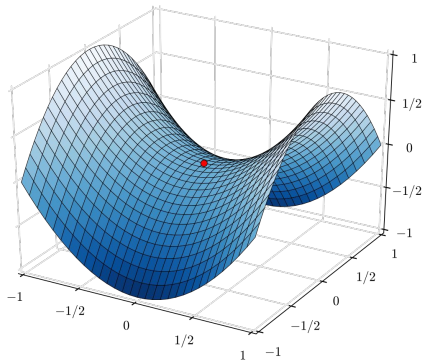
Question 3 (Information Asymmetry): What if the hospitals know more about patients than the agencies?

Key results: Regret exists even without confounding. Success depends on the degree of unmeasured heterogeneity.

Open questions and ongoing work

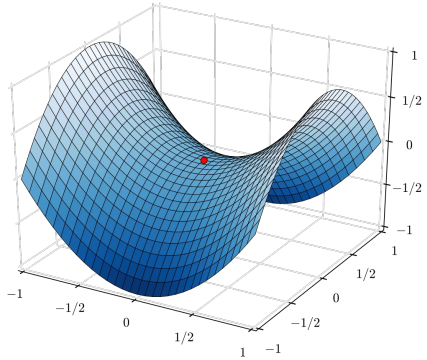
Open questions and ongoing work

Can **robust policy learning** help mitigate problems of information asymmetry?



Open questions and ongoing work

Can **robust policy learning** help mitigate problems of information asymmetry?



Incentives for information sharing: who has information on how to improve metrics, and when would they share it?

Open questions and ongoing work

Model extensions:

- Cost of treatment/resource constraints
- Patient decisions
- Proxies for health risk
- Competition between hospitals
- Competition between ranking platforms
- Truthfulness
- Hospital investment in improvement
- Multiple metrics
- …etc.

Behind every algorithm is a metric.

Many thanks!

serenalwang@berkeley.edu
serenalwang.com

Appendix

Example data sources for estimating $\hat{\mu}_0(x)$

Source 1: Auxiliary/historical untreated data

Requirements: Ignorability: $Y_i(0), Y_i(1) \perp\!\!\!\perp T_i | X_i, \;\; P(T_i = 1 | X_i = x) < 1$

Advantages: Can come from untreated observational data.

Challenges: Confounding; distribution shift.

Source 2: Agent's untreated units

Requirements: The agent's treatment policy depends only on X_i.

Advantages: No distribution shift.

Challenges: Unable to guarantee positivity.

Central threat: Confounding

Change 1: reward benefit from treatment

Reward function 2: "Average Treatment Effect on the Treated" (ATT)

$$w_{\mathrm{ATT}}(\mathbf{Y}, \mathbf{T}^{\pi}, \mathbf{X}) = \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}_0(X_i))T_i^{\pi}}{\sum_{i=1}^{n} T_i^{\pi}}$$

Estimated untreated counterfactual

Agent's best response: **Treat only the patients with the highest expected treatment effect.**

$$\pi^{w_{\mathrm{ATT}}}(x) = \mathbb{1}(x \in \arg\max_x \tau(x) \text{ and } \tau(x) > 0)$$

Proposition 2: Regret is **at most the max utility:** $R(\pi^{w_{\mathrm{ATT}}}) \leq \max_{\pi \in \Pi} V(\pi)$

Today

Patients

Selection

✓ ✗

Hospital

Mortality rate

%

CMS

CENTERS FOR MEDICARE & MEDICAID SERVICES

Centers for Medicare and Medicaid Services (CMS)

The US government spent **$1.3 billion** on quality metric development over the last 10 years (Wadhera et al. 2020).

Wadhera, R. K., Figueroa, J. F., Maddox, K. E. J., Rosenbaum, L. S., Kazi, D. S., & Yeh, R. W. (2020). Quality measure development and associated spending by the Centers for Medicare & Medicaid Services. *JAMA*, 323(16), 1614-1616.

Patients

Selection

✓ ✗

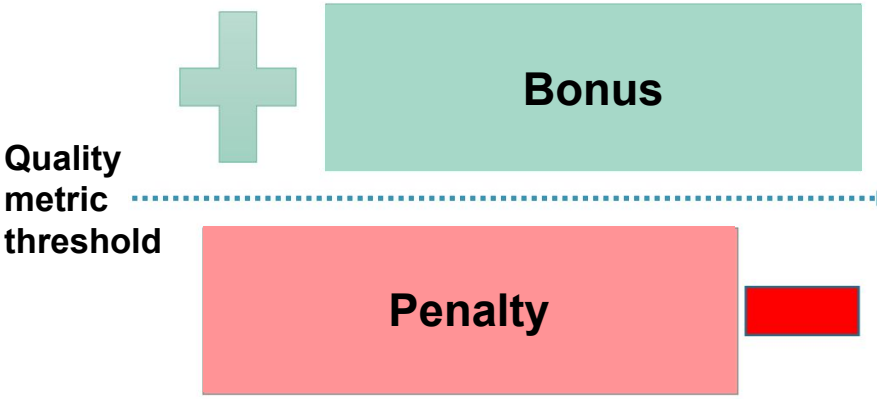
Hospital

Mortality rate

%

Centers for Medicare and Medicaid Services (CMS)

Merit-Based Incentive Payment System (MIPS)



Patients

Selection

✓ ✗

Hospital

Mortality rate

%

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Centers for Medicare and Medicaid Services (CMS)

Merit-Based Incentive Payment System (MIPS)

Quality metric threshold

Bonus

Penalty

Best Hospitals Honor Roll

Each year, U.S. News ranks hospitals in 15 specialties and 20 procedures and conditions. Here are the top 20 highest rated hospitals in the U.S.

2022-2023 Honor Roll Rankings

- #1 Mayo Clinic, Rochester, MN
- #2 Cedars-Sinai Medical Center, Los Angeles, CA
- #3 NYU Langone Hospitals, New York, NY

SEE FULL RANKINGS LIST »

Accounting for policy desiderata

Any positive function; policymaker's choice!

$$w^{g}_{\mathrm{TT}}(\mathbf{X}, \mathbf{T}^{\pi}, \mathbf{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_0(X_i)) T_i^{\pi} \, g(X_i)$$

Theorem 2 (Incentive Alignment): Reweighting preserves incentive-alignment.

Key result: this functional form **decouples** ranking desiderata from incentive alignment! A designer can optimize g for any ranking **policy desiderata**.

Example design choice: Boost for public vs. private institutions.

Bounds on regret under information asymmetry

Assumption (Bounded Unmeasured Heterogeneity): the effect of U on the untreated potential outcome given X is bounded:

$$E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$$

Theorem (Bounded Regret): the regret under reward w_{TT} is upper bounded by the degree of unmeasured heterogeneity: $R(\pi^{w_{\text{TT}}}) \leq 2\gamma_{\text{marg}}$

Low γ : X= smoking, U= drinking

"The magnitude of risk related to smoking is far larger than any ostensible benefit related to moderate drinking" (Mukamal 2006).

Mukamal, K. J. (2006). The effects of smoking and drinking on cardiovascular disease and risk factors. *Alcohol Research & Health*, 29(3), 199.

Bounds on regret under information asymmetry

Assumption (Bounded Unmeasured Heterogeneity): the effect of U on the untreated potential outcome given X is bounded:

$$E[|\mu_0(X_i) - \mu_0(X_i, U_i)|] \leq \gamma_{\text{marg}}$$

Theorem (Bounded Regret): the regret under reward w_{TT} is upper bounded by the degree of unmeasured heterogeneity: $R(\pi^{w_{\text{TT}}}) \leq 2\gamma_{\text{marg}}$

Low γ : X= smoking, U= drinking

"The magnitude of risk related to smoking is far larger than any ostensible benefit related to moderate drinking" (Mukamal 2006).

High γ : X = sex hormones, U = diabetes

"Cardiovascular risks associated with diabetes also appear to be higher in women" (Rodgers et al. 2019).

Rodgers, J. L. et al. (2019). Cardiovascular risks associated with gender and aging. *Journal of Cardiovascular Development and Disease,* 6(2), 19.

Mukamal, K. J. (2006). The effects of smoking and drinking on cardiovascular disease and risk factors. *Alcohol Research & Health*, 29(3), 199.