Maximizing Intervention Effectiveness

Vishal Gupta, Brian Rongqing Han, Song-Hee Kim, Hyung Paek（2017）

JESSIE GAO

Max intervention effectiveness (candidate group)

- Treatment: receive case management

- Causal effect: the number of ED visits for the next 6 months reduced by case management

- Goal: **decide K out of C candidates to apply treatment hoping to have max intervention effectiveness**

Practical background:

- Reduce ED utilization – treatment effect
- underpayment - reward
- Case management is expensive – need to select who to apply treatment

We seek to target at most $K > 0$ patients for intervention from a candidate population of size $C > K$ in order to maximize total intervention effectiveness. We adopt the Neyman-Rubin potential outcome framework for causal inference (Sekhon 2008). For each patient $c \in \{1, \ldots, C\}$, there exists a *fixed* tuple $(\mathbf{x}_c, y_c(0), y_c(1), r_c)$. The parameters $\mathbf{x}_c$ and r_c are assumed *known*, while $y_c(0)$ and $y_c(1)$ are *unknown* and represent potential outcomes. Specifically, $\mathbf{x}_c \in \mathcal{X}$ denotes pre-treatment covariates, such as demographic characteristics, of patient c, and may include both discrete and continuous variables.

causal effect of patient c as $\delta_c \equiv y_c(0) - y_c(1)$ the number of ED visits reduced

For each unit decrease in the outcome, we earn a reward of $r_c \geq 0$.

an estimate of the average charges per ED visit for patient c

intervention effectiveness of patient c to be $r_c \delta_c$

maximize the total intervention effectiveness as follows:

$$\max_{\mathbf{z} \in \mathcal{Z}} \sum_{c=1}^{C} z_c r_c \delta_c, \text{ where } \mathcal{Z} \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \leq K \right\}.$$

Available data
(study group)

* Evidence typically published in a research study doesn't include individual raw data

Table 1 Evidence of Causal Effects for Case Management from Shumway et al. (2008) for the Study Population

	Stratum 1 No. of ED Visits 5 - 11†	Stratum 2 No. of ED Visits ≥ 12	Total
No. of Patients			
Treatment	81 (32%)	86 (34%)	167 (66%)
Control	40 (15%)	45 (18%)	85 (34%)
No. of ED Visits in 6 Months			
Treatment, mean ± sd	2.5 ± 3.2	5.2 ± 5.6	3.9 ± 2.0*
Control, mean ± sd	4.6 ± 6.2	8.5 ± 9.6	6.7 ± 8.2*
CATE, mean ± sd, [95% CI]**	2.1 ± 1.0, [0.1, 4.1]	3.3 ± 1.6, [0.2, 6.4]	
SATE, mean ± sd, [95% CI]			2.7 ± 1.0‡, [0.8, 4.6]

Notes. † Patients are stratified based on number of ED visits in the previous year. * Approximated by taking the weighted average of the mean and variance for each group. For example, the mean outcome for the treatment group is $(81 \times 2.5 + 86 \times 5.2)/(81 + 86) = 3.9$ ED visits. ** CATE refers to the Conditional Average Treatment/Causal Effect within each stratum. We formally define CATE in Section 4.1. ‡ The standard deviation is approximated as $\sqrt{s_1^2/n_1 + s_2^2/n_2}$, where s_i is the standard deviation of the outcome and n_i is the number of people in group i for $i = 1, 2$.

Sample Average Treatment Effect (SATE).

$$\underline{I} \leq \frac{1}{S}\sum_{s=1}^{S} \delta^s \leq \overline{I}, \text{ where } \delta^s \equiv y^s(0) - y^s(1), \text{ for all } s \in \{1, \ldots, S\}.$$

(assuming we know the range of SATE, finite)

* Candidate group and study group are systematically different after detailed inclusion/exclusion criteria

Table 2 Summary Statistics for the Study and Candidate Populations

	Study Population ($S = 252$ patients)	Candidate Population ($C = 951$ patients)
Male	188 (75%)	42 (46%)
Race/Ethnicity		
African American	138 (54%)	75 (50%)
Hispanic	55 (22%)	7 (1%)
White	34 (13%)	83 (29%)
Other	28 (11%)	86 (20%)
Age, mean ± sd	43.3 ± 9.5	38.3 ± 12.5
No. of ED Visits in Previous Year*		
5 - 11	121 (48%)	860 (90%)
≥ 12	131 (52%)	91 (10%)
Most Frequent Diagnosis† during ED Visits	mental disorder (22%)	alcohol-related disorders (10%)
	injury (16%)	abdominal pain (6%)
	skin diseases (8%)	back problems (5%)
	endocrine disorders (5%)	nonspecific chest pain (4%)
	digestion disorders (5%)	connective tissue diseases (3%)
	respiratory illnesses (5%)	non-traumatic joint disorders (3%)

Notes. * Both populations only include patients who have had at least 5 ED visits. † Calculated from primary the ICD-10-CM diagnosis code using Clinical Classification Software (Elixhauser et al. 2014).

Max intervention effectiveness over CATE

Define Ψ(x): shared CATE for all x ∈ X w.r.t covariate

CATE of the study population $\mathbb{E}[\delta^{\tilde{s}}|\mathbf{x}^{\tilde{s}}=\mathbf{x}] \equiv \frac{1}{|\{s \mid \mathbf{x}^s=\mathbf{x}\}|}\sum_{s:\mathbf{x}^s=\mathbf{x}} \delta^s$

CATE of the candidate population is defined similarly.

$(\mathbf{x}^{\tilde{c}}, \delta^{\tilde{c}})$ for a patient $\tilde{c}$ picked uniformly at random from the candidate population.

ASSUMPTION 1 **(Shared CATEs)**. $\mathbb{E}[\delta^{\tilde{c}}|\mathbf{x}^{\tilde{c}}=\mathbf{x}] = \mathbb{E}[\delta^{\tilde{s}}|\mathbf{x}^{\tilde{s}}=\mathbf{x}], \quad \forall \mathbf{x} \in \mathcal{X}.$

$\Psi(\mathbf{x}) \equiv \mathbb{E}[\delta^{\tilde{c}}|\mathbf{x}^{\tilde{c}}=\mathbf{x}] = \mathbb{E}[\delta^{\tilde{s}}|\mathbf{x}^{\tilde{s}}=\mathbf{x}]$ decompose $\delta^{\tilde{c}} = \Psi(\mathbf{x}^{\tilde{c}}) + \epsilon^{\tilde{c}}$ $\mathbb{E}[\epsilon^{\tilde{c}}|\mathbf{x}^{\tilde{c}}]=0$ by construction

To motivate our robust approach, let $z_1, \ldots, z_C$ be any targeting policy that depends only on r_c and $\mathbf{x}_c$ for all $c \in \{1, \ldots, C\}$. The average intervention effectiveness of this policy is

$$\begin{aligned}\mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} \delta^{\tilde{c}}] &= \mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} \Psi(\mathbf{x}^{\tilde{c}})] + \mathbb{E}\left[\mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} \epsilon^{\tilde{c}} | \mathbf{x}^{\tilde{c}}]\right] \\ &= \mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} \Psi(\mathbf{x}^{\tilde{c}})] + \mathbb{E}\left[\mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} | \mathbf{x}^{\tilde{c}}] \mathbb{E}[\epsilon^{\tilde{c}} | \mathbf{x}^{\tilde{c}}]\right] \\ &= \mathbb{E}[z^{\tilde{c}} r^{\tilde{c}} \Psi(\mathbf{x}^{\tilde{c}})],\end{aligned}$$

ASSUMPTION 2 **(Conditionally Independent Rewards)**. *The rewards and idiosyncratic effects are conditionally independent given the observed covariates, i.e.,* $r^{\tilde{c}} \perp\!\!\!\perp \epsilon^{\tilde{c}} \mid \mathbf{x}^{\tilde{c}}$.

$$\max_{\mathbf{z} \in \mathcal{Z}} \sum_{c=1}^{C} z_c r_c \delta_c, \text{ where } \mathcal{Z} \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \leq K \right\}.$$ → (1) reduces to $\sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c)$.

What is Ψ():
assuming the format w.r.t covariate

$$\max_{\mathbf{z}\in\mathcal{Z}} \sum_{c=1}^{C} z_c r_c \delta_c, \text{ where } \mathcal{Z} \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \leq K \right\}.$$

$$\sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c)$$

- **Each individual**

- **CATE**

$$\Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x})$$

a parametric* class of functions

— Easy to verify $\mathbb{E}[\Psi(\mathbf{x}^s)] \in [\underline{I}, \bar{I}]$,

$$\underline{I} \leq \frac{1}{S}\sum_{s=1}^{S} \delta^s \leq \bar{I}, \text{ where } \delta^s \equiv y^s(0) - y^s(1), \text{ for all } s \in \{1,\ldots,S\}.$$

a linear combination of the description functions of covariates

Summary Statistics for covariates

Partition Description Functions: When there exists a natural partition of $\mathcal{X} = \bigcup_{i=1}^{I} \mathcal{X}_i$, e.g., patient race, studies often report the proportion of *type* i patients $\mu_i = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\mathbf{x}^s \in \mathcal{X}_i)$ for $i = 1, \ldots, I$. We model these statistics with description functions $\phi_i(\mathbf{x}) \equiv \mathbb{I}(\mathbf{x} \in \mathcal{X}_i)$ for $i = 1, \ldots, I-1$. Note that since $\mu_I = 1 - \sum_{i=1}^{I-1} \mu_i$ and $\phi_I(\mathbf{x}) = 1 - \sum_{i=1}^{I-1} \phi_i(\mathbf{x})$, it suffices to only specify these $I-1$ description functions to capture all I statistics. We assume for simplicity that $\mu_i > 0$ for $i = 1, \ldots, I$.

Linear Description Functions: When $\mathcal{X} \subseteq \Re^I$ contains continuous variables, studies often report their mean values in the study population $\mu_i = \frac{1}{S}\sum_{s=1}^{S} x_i^s$ for all $i = 1, \ldots, I$. We model these statistics with description functions $\phi_i(\mathbf{x}) \equiv x_i$ for $i = 1, \ldots, I$.

Quadratic Description Functions: When $\mathcal{X} \subseteq \Re^I$, studies may report, in addition to the mean $m_i \equiv \frac{1}{S}\sum_{s=1}^{S} x_i^s$, the standard deviation $\sigma_i^2 \equiv \frac{1}{S}\sum_{s=1}^{S} (x_i^s - m_i)^2$ of each covariate for all $i = 1, \ldots, I$. We model the mean m_i with the I description functions above and the standard deviation with additional I description functions: $\phi_{I+i}(\mathbf{x}) \equiv x_i^2$ and $\mu_{I+i} = m_i^2 + \sigma_i^2$ for all $i = 1, \ldots, I$.

$$\Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x})$$

$$\mu_g \equiv \frac{1}{S}\sum_{s=1}^{S} \phi_g(\mathbf{x}^s).$$

$$\mathbb{E}[\Psi(\mathbf{x}^{\tilde{s}})] = \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g$$

i-th dimension of covariate

Question (Saeedeh)

How do they obtain the probability distribution of the covariates (demographics)? Clinical trials usually only report mean and standard deviation of each covariate.

Table 2 **Summary Statistics for the Study and Candidate Populations**

	Study Population ($S = 252$ patients)	Candidate Population ($C = 951$ patients)
Male	188 (75%)	442 (46%)
Race/Ethnicity		
African American	138 (54%)	475 (50%)
Hispanic	55 (22%)	7 (1%)
White	34 (13%)	283 (29%)
Other	28 (11%)	186 (20%)
Age, mean ± sd	43.3 ± 9.5	38.3 ± 12.5
No. of ED Visits in Previous Year*		
5 - 11	121 (48%)	860 (90%)
≥ 12	131 (52%)	91 (10%)
Most Frequent Diagnosis† during ED Visits	mental disorder (22%)	alcohol-related disorders (10%)
	injury (16%)	abdominal pain (6%)
	skin diseases (8%)	back problems (5%)
	endocrine discorders (5%)	nonspecific chest pain (4%)
	digestion disorders (5%)	connective tissue diseases (3%)
	respiratory illnesses (5%)	non-traumatic joint disorders (3%)

Notes. * Both populations only include patients who have had at least 5 ED visits. † Calculated from primary the ICD-10-CM diagnosis code using Clinical Classification Software (Elixhauser et al. 2014).

Max worst-case intervention

$$\max_{\mathbf{z}\in\mathcal{Z}} \sum_{c=1}^{C} z_c r_c \delta_c, \text{ where } \mathcal{Z} \equiv \left\{ \mathbf{z} \in \{0,1\}^C \;\middle|\; \sum_{c=1}^{C} z_c \leq K \right\}.$$

$\sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c)$

- Each individual

- CATE

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi(\cdot)\in\mathcal{U}} \sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c).$$

- Worst-case (β_0^*, β^*)

$\Psi(\cdot)$ in an uncertainty set $\mathcal{U}$ restrict $\mathcal{U}$ to a parametric class of functions

$$\mathcal{U}_{\hat{\Gamma}} = \left\{ \Psi : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}), \;\; \underline{I} \leq \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \leq \bar{I}, \;\; \|\boldsymbol{\beta}\| \leq \hat{\Gamma} \right\}.$$

degree of heterogeneous causal effects in $\mathcal{U}$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_G)^T$.

Robust counterpart

THEOREM 3 **(Robust Counterpart)**.

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi(\cdot)\in\mathcal{U}} \sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c). \qquad \mathcal{U}_{\hat{\Gamma}} = \left\{ \Psi : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}),\ \ \underline{I} \le \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \le \bar{I},\ \ \|\boldsymbol{\beta}\| \le \hat{\Gamma} \right\}.$$

is equivalent to

$$\max_{\mathbf{z}\in\mathcal{Z}} \ \underline{I} \sum_{c=1}^{C} z_c r_c - \hat{\Gamma} \left\| \left(\sum_{c=1}^{C} z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|_*,$$

where $\|\cdot\|_*$ *is the dual norm to* $\|\cdot\|$.

Let $\|x\|$ *be a norm, then its dual norm* $\|x\|_*$ *is*

$$\|x\|_* = \max_{\|z\|\le 1} z^T x.$$

weighted ℓ_1 or ℓ_∞-norm,	a mixed-binary linear optimization problem
weighted ℓ_2-norm	a mixed-binary quadratic

NP-complete (nondeterministic polynomial time) but can be solved efficiently using off-the-shelf softwares

Covariate matching as regularization

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \underline{I}\sum_{c=1}^{C} z_c r_c - \hat{\Gamma}\left\|\left(\sum_{c=1}^{C} z_c r_c(\phi_g(\mathbf{x}_c)-\mu_g)\right)_{g=1}^{G}\right\|_*, \qquad \text{radius } \hat{\Gamma} \text{ controls the trade-off}$$

equivalent to approximating the targeted subpopulation's average effectiveness by the study population's average effectiveness (lower-bound/worst-case estimation of SATE in study group) minus a penalty that depends on the difference in demographic distributions between the targeted subpopulation and the study population.

larger the difference between the covariates in the study population and targeted patients is, the less reliable the SATE of the study population is as an estimate of causal effects in the targeted patients

Penalty coincides with common techniques used for covariate matching in causal inference. We can recover common matching procedures in special cases with appropriately chosen norms

- Chi-Square Matching under Partition Description Functions
- Mean Matching under Linear Description Functions

Example:
Mean Matching under Linear Description Functions

COROLLARY 3 (**Mean Matching under Linear Description Functions**). *Suppose* $\mathcal{X} \in \Re^G$, *and* $\phi_g(\mathbf{x}) = x_g$ *are our description functions with statistics* μ_g *for all* $g = 1, \ldots, G$. *For any positive definite matrix* $\mathbf{V} \in \Re^{G \times G}$, *consider uncertainty set* (7) *with a weighted* ℓ_2*-norm* $\|\cdot\|_V$. *Then, the robust targeting problem* (5) *is equivalent to*

define $\|\mathbf{t}\|_\mathbf{A} \equiv \sqrt{\mathbf{t}^T \mathbf{A} \mathbf{t}}$.

yield its dual norm: $\|\cdot\|_\mathbf{A}$-

$$\max_{\mathbf{z} \in \mathcal{Z}} \quad \underline{I} \sum_{c=1}^{C} z_c r_c - \hat{\Gamma} \left\| \sum_{c=1}^{C} z_c r_c (\mathbf{x}_c - \boldsymbol{\mu}) \right\|_{\mathbf{V}^{-1}}. \tag{11}$$

If we take **V** to be the covariance matrix of $\mathbf{x}^{\tilde{s}}$, we recognize the penalty term as the Mahalanobis distance, a common norm used in covariate matching techniques (see, e.g., Morgan et al. 2012). Alternatively, if we take **V** to be the diagonal with the variance of $\mathbf{x}^{\tilde{s}}$ as its entries, we recognize the penalty term as the mean matching penalty (see, e.g., Kallus 2016).

Question (Saeedeh)

Also some covariates may be correlated with each other. Does their method account for that?

Table 2 Summary Statistics for the Study and Candidate Populations

	Study Population ($S = 252$ patients)	Candidate Population ($C = 951$ patients)
Male	188 (75%)	442 (46%)
Race/Ethnicity		
African American	138 (54%)	475 (50%)
Hispanic	55 (22%)	7 (1%)
White	34 (13%)	283 (29%)
Other	28 (11%)	186 (20%)
Age, mean ± sd	43.3 ± 9.5	38.3 ± 12.5
No. of ED Visits in Previous Year*		
5 - 11	121 (48%)	860 (90%)
≥ 12	131 (52%)	91 (10%)
Most Frequent Diagnosis† during ED Visits	mental disorder (22%)	alcohol-related disorders (10%)
	injury (16%)	abdominal pain (6%)
	skin diseases (8%)	back problems (5%)
	endocrine discorders (5%)	nonspecific chest pain (4%)
	digestion disorders (5%)	connective tissue diseases (3%)
	respiratory illnesses (5%)	non-traumatic joint disorders (3%)

Notes. * Both populations only include patients who have had at least 5 ED visits. † Calculated from primary the ICD-10-CM diagnosis code using Clinical Classification Software (Elixhauser et al. 2014).

$$\Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x})$$

Later optimization stage, limiting norm of parameters as constraint?

$$\mathcal{U}_{\hat{\Gamma}} = \left\{ \Psi : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}),\;\; \underline{I} \le \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \le \bar{I},\;\; \|\boldsymbol{\beta}\| \le \hat{\Gamma} \right\}.$$

Performance Guarantee (1/2)
Worst-case performance is bounded by a constant

Worst-case performance is bounded by a constant that depends on the class of models and the true heterogeneous effect (Error from how linear the true CATEs in the description functions is as the assumption made at first stage to derive for format of CATE)…

$$(\beta_0^*, \boldsymbol{\beta}^*) \equiv \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{C} \sum_{c=1}^{C} \left(\Psi(\mathbf{x}_c) - \beta_0 - \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}_c) \right)^2$$

$$\max_{\mathbf{z} \in \mathcal{Z}} \; \underline{I} \sum_{c=1}^{C} z_c r_c - \hat{\Gamma} \left\| \left(\sum_{c=1}^{C} z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|_*,$$

be the best linear approximation to $\Psi(\cdot)$, and let $\varepsilon^2 \equiv \frac{1}{C} \sum_{c=1}^{C} \left(\Psi(\mathbf{x}_c) - \beta_o^* - \sum_{g=1}^{G} \beta_g^* \phi_g(\mathbf{x}_c) \right)^2$ be the average residual error. Note that $(\beta_0^*, \boldsymbol{\beta}^*)$ are unknown, although they do exist.

THEOREM 4 **(Worst-Case Performance of Robust Targeting)**. *Let* $\mathbf{z}^{Rob}$ *be an optimizer of problem* (8). *For any uncertainty* $\mathcal{U}_{\hat{\Gamma}}$ *of the form in* (7), *if* $(\beta_0^*, \boldsymbol{\beta}^*) \in \mathcal{U}_{\hat{\Gamma}}$, *then*

$$\sum_{c=1}^{C} z_c^{Rob} r_c \Psi(\mathbf{x}_c) \geq -\sqrt{C}\varepsilon \| (z_c^{Rob} \cdot r_c)_{c=1}^{C} \| \geq -\sqrt{C}\varepsilon K \max\{r_c\}_{c=1}^{C}.$$

Theorem 4 asserts that if $\hat{\Gamma}$ is large enough that $\mathcal{U}_{\hat{\Gamma}}$ contains $(\beta_0^*, \boldsymbol{\beta}^*)$, the expected performance of the robust model is bounded below by a constant that depends on the extent to which the CATE is linear over the description functions. It does not require $\Psi(\cdot)$ to be linear in the description

Performance Guarantee (2/2)
never worse than not targeting

… And under some mild assumptions, this constant is zero, ensuring that our robust approach is never worse than not targeting, even when the treatment could be potentially harmful ！ In special cases where the true CATE is known to be linear, such as when |X | is finite, as in Remark 5, this constant is 0。

$$\varepsilon^2 \equiv \frac{1}{C}\sum_{c=1}^{C}\left(\Psi(\mathbf{x}_c) - \beta_o^* - \sum_{g=1}^{G}\beta_g^*\phi_g(\mathbf{x}_c)\right)^2$$

REMARK 5. When $|\mathcal{X}|$ is finite and the partition consists of singletons, *every* CATE $\Psi(\mathbf{x})$ can be written in the form $\Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G}\beta_g\mathbb{I}(\mathbf{x}\in\mathcal{X}_g)$ for some $(\beta_0,\boldsymbol{\beta})\in\Re^{G+1}$ and $\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}} = \sqrt{\mathrm{Var}(\Psi(\mathbf{x}^{\tilde{s}}))}$. We consider this the canonical setting for Corollary 2.

We stress that in (7), only the choice of norm and radius $\hat{\Gamma}$ are user-specified; all other parameters are determined by the study evidence. For the specific choice $\|\boldsymbol{\beta}\|^2 \equiv \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}$, where $\Sigma_{gg'} \equiv \mathbb{E}[(\phi_g(\mathbf{x}^{\tilde{s}}) - \mu_g)(\phi_{g'}(\mathbf{x}^{\tilde{s}}) - \mu_{g'})]$ for all $g,g' = 1,\ldots,G$, the constraint $\|\boldsymbol{\beta}\| \le \hat{\Gamma}$ is equivalent to $\mathrm{Var}(\Psi(\mathbf{x}^{\tilde{s}})) \le \hat{\Gamma}^2$ if $\Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G}\beta_g\phi_g(\mathbf{x})$. We use this insight to motivate specific choices of norm in special cases in what follows.

Scoring Rules Method

approximate the unknown causal effect δ_c with some observable proxy $\hat{\delta}_c$ and then rank patients accordingly.

Constant Effect Sizes and Reward Scoring: If one believes that the true causal effect is constant, i.e., $\delta_c = \delta_0 > 0$ for all $c \in \{1, \ldots, C\}$, then no matter what the value of δ_0 is, using the proxy $\hat{\delta}_c = 1$ and ranking patients by r_c (i.e., reward scoring, or r-scoring) yields an optimal solution to problem (1). The assumption of constant effect sizes is common in statistical inference for randomized control trials. In particular, the most common approach for estimating the sampling variance of the SATE estimator assumes that the causal effects are constant across all individuals in the study population (Imbens 2004).

$$(\hat{\delta}_c = 1)$$

Proportional Effect Sizes and Outcome Scoring: If one believes that the true causal effect is proportional to the outcome without treatment, i.e., $\delta_c = \alpha y_c(0)$ for all $c \in \{1, \ldots, C\}$ and some $\alpha > 0$, then no matter what the value of α is, using the proxy $\hat{\delta} = y_c(0)$ and ranking patients by $r_c y_c(0)$ (i.e., outcome scoring, or $ry(0)$-scoring) yields an optimal solution to problem (1).

$$\text{proxy } \hat{\delta} = y_c(0)$$

In words, outcome scoring targets *high-risk* patients. Many studies have developed statistical or machine-learning models to predict the number of ED visits by a specific patient in the near future (Billings and Raven 2013). They estimate $y_c(0)$ (i.e., number of ED visits for patient c) and suggest focusing on patients with large estimates. Implicitly, such recommendations assume that the true causal effect δ_c is proportional to $y_c(0)$.

Sufficient conditions for their optimality

When three is no harmful effect

tight performance bounds for current practice (scoring rules).

THEOREM 1 (**Worst-Case Performance of Scoring Rules with Benign Treatment**). *Without loss of generality, index patients so that* $r_1\hat{\delta}_1 \geq \cdots \geq r_C\hat{\delta}_C \geq 0$. *Suppose* $K \leq C/2$ *and there exists* $0 < \underline{\delta} < \overline{\delta} < \infty$ *such that* $\delta_c/\hat{\delta}_c \geq \underline{\delta} > 0$, *for all* $c \in \{1,\ldots,K\}$, *and* $\delta_c/\hat{\delta}_c \leq \overline{\delta}$, *where* $\overline{\delta} > 0$ *for all* $c \in \{K+1,\ldots,C\}$. *Then, the* $r\hat{\delta}$*-scoring rule obtains at least* $\omega(\underline{\delta}/\overline{\delta})$ *of the full-information benchmark optimal value, where*

$$\omega(\underline{\delta}/\overline{\delta}) \equiv \frac{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{K} r_c\hat{\delta}_c}{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{k^*} r_c\hat{\delta}_c + \sum_{c=K+1}^{2K-k^*} r_c\hat{\delta}_c} \tag{2}$$

and

$$k^* = \begin{cases} 0, & \text{if } (\underline{\delta}/\overline{\delta}) \leq r_{2K}\hat{\delta}_{2K}/r_1\hat{\delta}_1 \\ \arg\max\{c \mid 1 \leq c \leq K,\ (\overline{\delta}/\underline{\delta}) \geq r_{2K-c+1}\hat{\delta}_{2K-c+1}/r_c\hat{\delta}_c\}, & \text{otherwise.} \end{cases}$$

that $\delta_c/\hat{\delta}_c \geq \underline{\delta} > 0$, for all $c \in \{1,\ldots,K\}$. Indeed, since $r_c \geq 0$, this implies that $r_c\delta_c \geq 0$ for all $c \in \{1,\ldots,K\}$, i.e., targeted patients do not experience adverse effects.

Sufficient conditions for their optimality

When three is no harmful effect

"Intuitively, scoring rules should improve as the degree of correspondence increases. We formalize theses intuition by proving that w(·) shares these structural features: "

$$\omega(\underline{\delta}/\overline{\delta}) \equiv \frac{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{K} r_c\hat{\delta}_c}{(\underline{\delta}/\overline{\delta})\sum_{c=1}^{k^*} r_c\hat{\delta}_c + \sum_{c=K+1}^{2K-k^*} r_c\hat{\delta}_c}$$

only through the ratio $\underline{\delta}/\overline{\delta}$. I

COROLLARY 1. *Under the assumptions of Theorem 1,*

(1) $\omega(\underline{\delta}/\overline{\delta})$ *is increasing in* $\underline{\delta}/\overline{\delta}$; When the degree of heterogeneity is 0, reward scoring is optimal

(2) If $\underline{\delta}/\overline{\delta} \geq r_{K+1}\hat{\delta}_{K+1}/r_K\hat{\delta}_K$, $\omega(\underline{\delta}/\overline{\delta}) = 1$ *and the* $r\hat{\delta}$*-scoring rule is optimal;* Full-information benchmark optimal value

(3) $\omega(\underline{\delta}/\overline{\delta}) \to 0$ *as* $\underline{\delta}/\overline{\delta} \to 0$.

this ratio measures the degree of correspondence between the proxy and the true causal effect

Intuitively, $\overline{\delta}$ measures how much we may have underestimated the causal effects of patients that were not picked, while $\underline{\delta}$ measures how much we may have overestimated the causal effect of patients that were picked. With this interpretation, the critical assumption in Theorem 1 is

$\underline{\delta}_c/\hat{\delta}_c \geq \underline{\delta} > 0$, *for all* $c \in \{1,\ldots,K\}$,

$\overline{\delta}_c/\hat{\delta}_c \leq \overline{\delta}$, *where* $\overline{\delta} > 0$ *for all* $c \in \{K+1,\ldots,C\}$

Problem of Scoring Rules

proxy vs true real effect

Without loss of generality, index patients so that $r_1\hat{\delta}_1 \geq \cdots \geq r_C\hat{\delta}_C \geq 0$.

REMARK 1 (SCORING RULES CAN BE WORSE THAN NOT TARGETING). Index patients as in Theorem 1, and suppose that $\delta_c/\hat{\delta}_c = \underline{\delta} < 0$ for all $c \in \{1, \cdots, K\}$ and $\sum_{c\in B^*} r_c\delta_c > 0$. Such a scenario might occur if the treatment only benefitted a small subgroup of the population but potentially harmed others, and scoring rules cannot perfectly determine which is which. Then, $r\hat{\delta}$-scoring has intervention effectiveness $\underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c < 0$, which, in an absolute sense, is worse than not providing treatment to anyone. In terms of relative performance, the performance can be arbitrarily bad when the treatment is only marginally effective since

$$\frac{\underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c}{\sum_{c\in B^*} r_c\delta_c} \to -\infty \text{ as } \sum_{c\in B^*} r_c\delta_c \to 0.$$

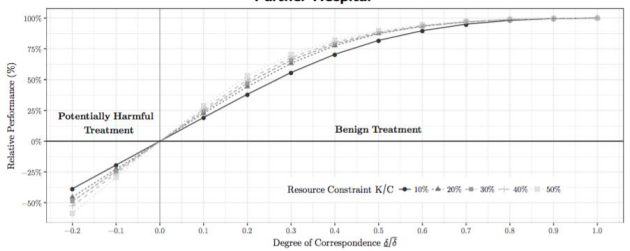
B is the best solution if we observe fully-observed casual effect

In terms of the performance difference, this may be as large as

$$\overline{\delta}\sum_{c=K+1}^{2K} r_c\hat{\delta}_c - \underline{\delta}\sum_{c=1}^{K} r_c\hat{\delta}_c,$$

if $B^* = \{K+1, \ldots, 2K\}$ and $\delta_c/\hat{\delta}_c = \overline{\delta} > 0$ for all $c \in B^*$. Such a scenario might occur if there do exist high-reward patients who would benefit from the treatment, but the particular scoring rule does not identify them.

Figure 1 Worst-Case Relative Performance of Reward Scoring (r-Scoring) for Case Management in Our Partner Hospital



Note. When the treatment is benign, we plot the worst-case relative performance bound (2) provided in Theorem 1. When the treatment is potentially harmful, the worst-case relative performance is $-\infty$, as mentioned in Remark 1. Thus, we plot $\underline{\delta}\sum_{c=1}^{K} r_c/\overline{\delta}\sum_{c=K+1}^{2K} r_c$ for comparison.

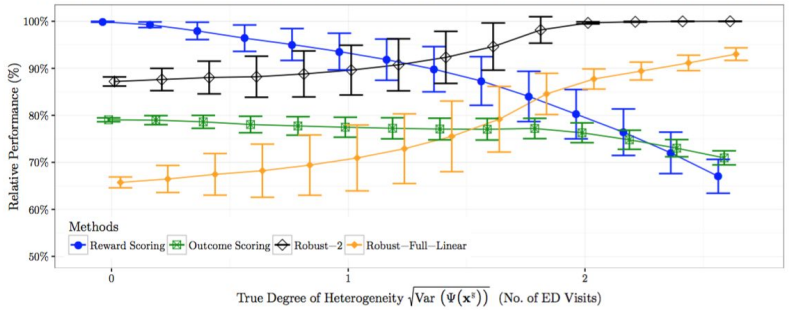
when case management may be ineffective or increase ED visits (i.e., potentially harmful treatment)

Intuition (later proved by experiment) long tail o reward: Intuitively, when there exists a small proportion of patients with very high marginal rewards, targeting these patients is “risky.”

Robust optimization vs Scoring Rules

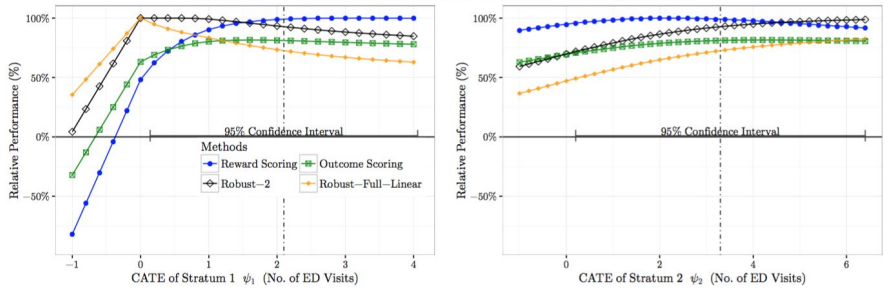
Result of Section 5.4 (assuming strata based on previous ED visit alone capturing all heterogeneity in casual effect

Figure 2 Relative Performance as Both ψ_1 and ψ_2 Vary Uniformly within Their Confidence Intervals



Note. For a fixed value of $\sqrt{\mathrm{Var}(\Psi(\mathbf{x}^s))}$, we summarize the relative performance using its mean (points) and standard deviation (error bars) for each method.

Figure 3 Relative Performance as Either ψ_1 or ψ_2 Varies



Note. In the left panel, we fix $\psi_2 = 3.3$ at its point estimate, varying only ψ_1, while in the right panel, we fix $\psi_1 = 2.1$ at its point estimate, varying only ψ_2. The dashed vertical lines correspond to the point estimates of the CATEs, which vary.

Using data from our partner hospital, we show that

- our robust approach performs almost as well as scoring rules when the degree of heterogeneity in causal effects is small and,
- can perform much better than scoring rules as the degree of heterogeneity increases, especially when the treatment is potentially harmful.

Robust optimization vs Scoring Rules

Scoring rules can be special cases of robust optimization.

$$\max_{\mathbf{z}\in\mathcal{Z}} \quad \underline{I}\sum_{c=1}^{C} z_c r_c - \hat{\Gamma}\left\|\left(\sum_{c=1}^{C} z_c r_c(\phi_g(\mathbf{x}_c)-\mu_g)\right)_{g=1}^{G}\right\|_*, \qquad \hat{\Gamma}=0, \quad \text{reduces to reward scoring.}$$

Table 4 Characteristics and Reward-Weighted Average Covariates for the Targeted Patients by Method

	Study Population	Candidate Population	Reward Scoring	Outcome Scoring	Robust-2	Robust-Full-Linear
Characteristics of Targeted Patients						
Avg. Charges Per ED Visit in 2014 r_c ($)		3,324	7,547	5,965	6,570	4,751
Avg. No. of ED Visits from 1/1/2015 to 6/30/2015 $y_c(0)$		2.3	2.2	5.3	3.4	3.5
Weighted Avg. Pre-Treatment Covariates*						
Demographics						
Male	75%	48%	51%	52%	54%	75%
African American	54%	46%	43%	44%	41%	54%
Hispanic	22%	0.5%	0.3%	0.1%	0.1%	2%
White	13%	34%	41%	42%	43%	16%
Age	43.3	40.7	44.9	44.9	45.3	43.3
Two Strata from Shumway et al. (2008)						
Stratum 1: 5 - 11 ED Visits in Previous Year	48%	90%	90%	83%	76%	67%

Note. * Except for the study population, we show the reward-weighted average of pre-treatment covariates. Thus, the reward-weighted summary statistics for the candidate population are different from those in Table 3.

- **Robust-2:** We solve problem (10) using a partition description function for str

lently, we add the constraints $\beta_g = 0$ to all other description functions.) The correspond

statistics, i.e., the proportion of patients in each stratum, are given by Table 1.

Because of their different choices of description functions, the two robust methods match the distribution of covariates in the study population differently. Robust-2 attempts to match the proportion of patients in each stratum. Specifically, Table 4 shows that there are only 48% patients in stratum 1 in the study population, in contrast to 90% of the reward-weighted proportion of patients in stratum 1 in the candidate population. Thus, Robust-2 targets proportionally fewer stratum 1 patients, yielding an overall percentage of 76%. Notice that although this proportion is closer to the study population's 48%, it is not an exact match since the robust method also balances the competing objective of targeting higher-reward patients (recall Remark 3). In our candidate population, many stratum 2 patients have much lower r_c than typical stratum 1 patients (Figure EC.2 in the e-companion). Completely matching the proportion of stratum 1 patients would entail a significant loss in rewards. Notice also that although Robust-2 improves the matching in the proportion of patients in each stratum, it exacerbates differences in other covariates, such as the proportion of Hispanic patients.

- **Robust-Full-Linear:** We solve problem (11) using partition description functions for strata, gender, and race and a linear description function for age. The corresponding summary statistics, i.e., the mean and standard deviation of each covariate, are given by Table 2.

Similar observations can be made about Robust-Full-Linear. It more closely matches the means of the covariates in the study population than in the candidate population but does not always achieve exact matching. For example, our candidate population has very few Hispanic patients, so Robust-Full-Linear is unable to fully match the 22% of Hispanic patients in the study population. For all other covariates, it achieves a reasonably close match.

Question (Nirvan)

How does the "covariate matching as regularization penalty" grow as difference in subpopulations grows? I imagine it would be hard for this not to grow quickly if you are restricting yourself to just summary statistics. Whereas other individual level approaches should do much better?

Question (Angela)

Question for Maximizing Intervention Effectiveness: re: the covariate matching as regularization assumption: Intuitively, if we interpret the regularization term as covariate matching to the moments of the study population: would we first remove low-reward individuals which were outlying on high-variance covariates in the study population from the targeting set?

Question (Amu4)

One of the areas where finding CATE is really useful is when the treatment can have negative effects on some groups. However, this paper is using published studies, which often suffer from often sampling (which subjects are most convenient for the researcher) and reporting (pressure find significant results, non reporting of null results) biases. I think this would make it hard to estimate CATE in the instances where it really matters; is there a way to account for this censoring to correct estimates?

Sensitivity analysis

Long tail of reward

K/C

etc

Other advantages（1/2）

flexible enough to easily accommodate side constraints on the targeting, such as budget, operational or fairness constraints.

THEOREM 3 **(Robust Counterpart)**.

$$\max_{\mathbf{z}\in\mathcal{Z}} \min_{\Psi(\cdot)\in\mathcal{U}} \sum_{c=1}^{C} z_c r_c \Psi(\mathbf{x}_c). \qquad \mathcal{U}_{\hat{\Gamma}} = \left\{ \Psi : \mathcal{X} \mapsto \mathbb{R} \;\middle|\; \Psi(\mathbf{x}) = \beta_0 + \sum_{g=1}^{G} \beta_g \phi_g(\mathbf{x}),\ \ \underline{I} \le \beta_0 + \sum_{g=1}^{G} \beta_g \mu_g \le \bar{I},\ \ \|\boldsymbol{\beta}\| \le \hat{\Gamma} \right\}.$$

is equivalent to

$$\max_{\mathbf{z}\in\mathcal{Z}} \ \underline{I} \sum_{c=1}^{C} z_c r_c - \hat{\Gamma} \left\| \left(\sum_{c=1}^{C} z_c r_c (\phi_g(\mathbf{x}_c) - \mu_g) \right)_{g=1}^{G} \right\|_*, \qquad \textit{where } \|\cdot\|_* \textit{ is the dual norm to } \|\cdot\|.$$

Let $\|x\|$ be a norm, then its dual norm $\|x\|_*$ is

$$\|x\|_* = \max_{\|z\|\le 1} z^T x.$$

weighted ℓ_1 or ℓ_∞-norm,	a mixed-binary linear optimization problem
weighted ℓ_2-norm	a mixed-binary quadratic

Similarly, adding linear constraints on $\mathbf{z}$ does not significantly increase the complexity of the model. Thus, fairness constraints such as requiring that equal numbers of men and women be targeted for treatment can easily be incorporated.

As with most robust optimization models, including additional convex constraints on $\boldsymbol{\beta}$ in (7) does not significantly increase the complexity of the model. Thus, we might incorporate a priori knowledge of the structure on $\Psi(\cdot)$ by enforcing $\beta_g = 0$ for some g or by bounding its magnitude

Other advantages (2/2)

our work also connects to meta-regression techniques in statistics that aim to “pool” the results of different published studies

An interesting extension involves incorporating evidence from multiple papers. For concreteness, consider the case of J papers with corresponding interval $[\underline{I}^j, \overline{I}^j]$, description functions $\phi_g^j(\cdot)$ and statistics μ_g^j for all $g = 1, \ldots, G^j$ and $j = 1, \ldots, J$. A first attempt to generalize (7) might consider $\Psi(\cdot)$, which is a linear combination of the entire set of description functions $\{\phi_g^j(\cdot)\}_{g=1,\ldots,G^j, j=1,\ldots,J}$. Unfortunately, as in Section 4.1, for this rich a representation, it becomes difficult to verify that $\mathbb{E}[\Psi(\mathbf{x}^{\tilde{s}^j})] \in [\underline{I}^j, \overline{I}^j]$, where $\tilde{s}^j$ is a patient randomly drawn from study population j.

A practical compromise is to limit $\Psi(\cdot)$ to be a linear combination of description functions shared by all J papers. In this case, the constraints $\mathbb{E}[\Psi(\mathbf{x}^{\tilde{s}^j})] \in [\underline{I}^j, \overline{I}^j]$ reduce to linear constraints on the corresponding coefficients, and the usual techniques yield a robust counterpart. We argue that this approach is practically reasonable because most studies report the same basic demographic summary statistics for their study population, such as age, gender, and race. A full treatment of handling multiple papers with different description functions is beyond the scope of this paper.

4.4. Choosing $\hat{\Gamma}$

We next propose a heuristic for choosing $\hat{\Gamma}$. Inspired by Theorem 4, an ideal choice of $\hat{\Gamma}$ would be $\|\boldsymbol{\beta}^*\| \approx \sqrt{\text{Var}(\Psi(\mathbf{x}^{\tilde{s}}))}$. In applications where good bounds $(\underline{b}, \bar{b})$ such that $\underline{b} \leq \Psi(\mathbf{x}^{\tilde{s}}) \leq \bar{b}$ are available, Popoviciu's inequality on variance yields $\sqrt{\text{Var}(\Psi(\mathbf{x}^{\tilde{s}}))} \leq \frac{\bar{b}-\underline{b}}{2}$ (Bhatia and Davis 2000). In the absence of other information, $\sqrt{\text{Var}(\Psi(\mathbf{x}^{\tilde{s}}))}$ might be anywhere in the interval $[0, \frac{\bar{b}-\underline{b}}{2}]$. Consequently, we suggest choosing $\hat{\Gamma}$ to be the average $\frac{\bar{b}-\underline{b}}{4}$ over this interval. In applications where good bounds of $(\underline{b}, \bar{b})$ are unavailable, we recommend setting $(\underline{b}, \bar{b}) = (\underline{I}, \bar{I})$.

We choose $\hat{\Gamma}$ via the heuristic in Section 4.4, yielding $\hat{\Gamma} = 0.95$ for both robust methods. Finally, as observed by Iancu and Trichakis (2013), robust optimization problems frequently exhibit multiple optimal solutions, and some care should be taken in selecting a particular solution. In our case study, we use the Pareto robust optimal solution corresponding to the realization $\Psi(\cdot) = \underline{I}\mathbf{e}$, which can be computed using the techniques of Iancu and Trichakis (2013). Intuitively, this yields a robust solution that is non-dominated with respect to the model in which CATEs are homogeneous, i.e., all patients respond to the treatment identically.

Summary

Method	Scoring Rules	Robust optimization
		Use description functions and their statistics maximize the worst-case performance over an uncertainty set of models for the heterogeneous causal e↵ect that are consistent with the published study evidence Model: a parametric* class of functions
	sufficient conditions for their optimality and a tight performance guarantee when they are suboptimal	