Towards Reliable Inference for Precision Medicine

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

August 9, 2021 @ JSM

Manuscripts and slides are available at `http://statslab.cam.ac.uk/~qz280/`.

Precision medicine or Jenga?

This talk

Motivation

Statistical inference for precision medicine needs to be reliable.

Two attempts

1. Post-selection inference for effect modifiers.
   - Reference: Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. "Selective inference for effect modification via the lasso. *arXiv:1705.08020* (2017).
2. Sensitivity analysis for individualized treatment rules.
   - Reference: Bo Zhang, Jordan Weiss, Dylan S Small, Qingyuan Zhao. "Selecting and ranking individualized treatment rules with unmeasured confounding". *Journal of the American Statistical Association* 116.533 (2021): 295–308.

Outline

Post-selection inference for effect modifiers

Sensitivity analysis for individualized treatment rules

Motivations

Effect modification = Treatment effect varies across individuals.

- Equivalently, there is an interaction effect between treatment and covariates.
- Synonyms: heterogeneous treatment effect, subgroup analysis.

Why investigate effect modification?

- Optimal treatment regime (Murphy 2003; Kosorok and Laber 2019).
- Extrapolation of average causal effect to a different population (Stuart *et al.* 2011).
- Better understanding the causal mechanism (Grobbee and Hoes 2009; VanderWeele and Robins 2007).
- Make inference less sensitive to unmeasured confounding (Hsu *et al.* 2013).

Motivations

Effect modifiction = Treatment effect varies across individuals.

- Equivalently, there is an interaction effect between treatment and covariates.
- Synonyms: heterogeneous treatment effect, subgroup analysis.

Why investigate effect modification?

- Optimal treatment regime (Murphy 2003; Kosorok and Laber 2019).
- Extrapolation of average causal effect to a different population (Stuart *et al.* 2011).
- Better understanding the causal mechanism (Grobbee and Hoes 2009; VanderWeele and Robins 2007).
- Make inference less sensitive to unmeasured confounding (Hsu *et al.* 2013).

Existing methods

Classical methods: Subgroup analysis and regression analysis

- Prespecified subgroups/interactions:
  - **Strengths:** Free of selection bias; scientifically rigorous.
  - **Limitations:** Cannot test too many; no flexibility.
- Post hoc subgroups [Scheffé, Tukey (1950s)].
  - **Limitations:** Low power.
- Sample splitting: use part of the data for discovery and the other part for confirmation.
  - **Limitations:** Some information is discarded.

More recent methods (list is not complete)

- **Bayesian ensemble** (Hill 2011; Green and Kern 2010).
- **Outcome-weighted classification** (Zhao *et al.* 2012).
- **Lasso-regularized regression** (Qian and Murphy 2011; Imai and Ratkovic 2013; Tian *et al.* 2014).
- **Tree-based statistical learning** (Hsu *et al.* 2015; Athey *et al.* 2019; Powers *et al.* 2018).

Existing methods

Classical methods: Subgroup analysis and regression analysis

- Prespecified subgroups/interactions:
  - **Strengths:** Free of selection bias; scientifically rigorous.
  - **Limitations:** Cannot test too many; no flexibility.
- Post hoc subgroups [Scheffé, Tukey (1950s)].
  - **Limitations:** Low power.
- Sample splitting: use part of the data for discovery and the other part for confirmation.
  - **Limitations:** Some information is discarded.

More recent methods (list is not complete)

- **Bayesian ensemble** (Hill 2011; Green and Kern 2010).
- **Outcome-weighted classification** (Zhao *et al.* 2012).
- **Lasso-regularized regression** (Qian and Murphy 2011; Imai and Ratkovic 2013; Tian *et al.* 2014).
- **Tree-based statistical learning** (Hsu *et al.* 2015; Athey *et al.* 2019; Powers *et al.* 2018).

A General Setup

A nonparametric structural mean model:

$$\mathrm{E}[Y_i(t) \mid \boldsymbol{X}_i] = \eta(\boldsymbol{X}_i) + t \cdot \Delta(\boldsymbol{X}_i), \ \ i = 1, \ldots, n.$$

- $\Delta(\mathbf{x})$ is the parameter of interest.
- **Saturated** if treatment is binary, $t \in \{0, 1\}$.
- In this case, $\Delta(\boldsymbol{x}) = \mathrm{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$ is the **conditional average treatment effect (CATE)**.

Assumption (Standard assumptions for point identification)

(A) *Consistency of the observed outcome:* $Y_i = Y_i(T_i)$;

(B) *Unconfoundedness:* $Y_i(t) \perp\!\!\!\perp T_i \mid \mathbf{X}_i, \ \forall t \in \mathcal{T}$;

(C) *Positivity/Overlap:* $\mathrm{Var}(T_i \mid \mathbf{X}_i = \mathbf{x})$ *exists and is bounded away from* 0 *for all* $\mathbf{x}$.

Randomized experiments would be a special case for (B).

A General Setup

A nonparametric structural mean model:

$$\mathrm{E}[Y_i(t) \mid \boldsymbol{X}_i] = \eta(\boldsymbol{X}_i) + t \cdot \Delta(\boldsymbol{X}_i), \;\; i = 1, \ldots, n.$$

- $\Delta(\mathbf{x})$ is the parameter of interest.
- **Saturated** if treatment is binary, $t \in \{0, 1\}$.
- In this case, $\Delta(\boldsymbol{x}) = \mathrm{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$ is the **conditional average treatment effect (CATE)**.

Assumption (Standard assumptions for point identification)

(A) *Consistency of the observed outcome:* $Y_i = Y_i(T_i)$;

(B) *Unconfoundedness:* $Y_i(t) \perp\!\!\!\perp T_i \mid \mathbf{X}_i, \; \forall t \in \mathcal{T}$;

(C) *Positivity/Overlap:* $\mathrm{Var}(T_i \mid \mathbf{X}_i = \mathbf{x})$ *exists and is bounded away from* 0 *for all* $\mathbf{x}$.

Randomized experiments would be a special case for (B).

Prediction or inference?

Objective of most statistical methods: Accurate estimation/prediction

- CATE: $\Delta(\boldsymbol{x})$; or
- Optimal treatment rule: $I(\Delta(\boldsymbol{x}) > 0)$.

Objective in practice: Often less straightforward

Example: In 2017 Atlantic Causal Inference Conference, a workshop was organized to compare different approaches to investigate the effect modification. The organizers asked the following questions:

1. Was the (educational) intervention effective?
2. Two variables were hypothesized to modify the treatment effect. Are these hypotheses supported by empirical data?
3. Are there other effect modifiers?

Prediction or inference?

Objective of most statistical methods: Accurate estimation/prediction

- CATE: $\Delta(\boldsymbol{x})$; or
- Optimal treatment rule: $I(\Delta(\boldsymbol{x}) > 0)$.

Objective in practice: Often less straightforward

Example: In 2017 Atlantic Causal Inference Conference, a workshop was organized to compare different approaches to investigate the effect modification. The organizers asked the following questions:

1. Was the (educational) intervention effective?
2. Two variables were hypothesized to modify the treatment effect. Are these hypotheses supported by empirical data?
3. Are there other effect modifiers?

One solution: Post-selection inference

Tradeoff: Accuracy vs. Interpretability

	Univariate	**Selected submodel**	Full model	Machine learning
Model of $\Delta(\boldsymbol{x})$	$\alpha_j + x_j^T \beta_j$	$\alpha_{\hat{\mathcal{M}}} + \boldsymbol{x}_{\hat{\mathcal{M}}}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}}$	$\alpha + \boldsymbol{x}^T \boldsymbol{\beta}$	e.g. additive trees
Accuracy	Poor	**Good**	Good	Very good
Interpretability	Very good	**Good**	Poor	Very poor
Inference	Easy, but many false positives	**Need to consider model selection**	Semiparametric or high dim. theory	No clear objective

Overview of the method

1. **De-confounding:** Remove confounding bias without making parametric assumptions.
2. **Model selection:** Obtain a simple and reasonably good approximation of $\Delta(\boldsymbol{x})$.
3. **Post-selection inference:** Make statistical inference for the selected submodel.

One solution: Post-selection inference

Tradeoff: Accuracy vs. Interpretability

	Univariate	**Selected submodel**	Full model	Machine learning
Model of $\Delta(\boldsymbol{x})$	$\alpha_j + x_j^T \beta_j$	$\alpha_{\hat{\mathcal{M}}} + \boldsymbol{x}_{\hat{\mathcal{M}}}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}}$	$\alpha + \boldsymbol{x}^T \boldsymbol{\beta}$	e.g. additive trees
Accuracy	Poor	**Good**	Good	Very good
Interpretability	Very good	**Good**	Poor	Very poor
Inference	Easy, but many false positives	**Need to consider model selection**	Semiparametric or high dim. theory	No clear objective

Overview of the method

1. **De-confounding:** Remove confounding bias without making parametric assumptions.
2. **Model selection:** Obtain a simple and reasonably good approximation of $\Delta(\boldsymbol{x})$.
3. **Post-selection inference:** Make statistical inference for the selected submodel.

Background: Post-selection inference

Suppose we have noisy observations of Δ: $Y_i = \Delta(\mathbf{X}_i) + \epsilon_i,\ i = 1, \ldots, n$.

- Model selection procedure: $\{\mathbf{X}_i, Y_i\}_{i=1}^n \mapsto \hat{\mathcal{M}}$.
- **Conditional confidence interval**:

$$\mathrm{P}\Big(\big(\beta^*_{\mathcal{M}}\big)_j \in [D_j^-, D_j^+] \;\Big|\; \hat{\mathcal{M}} = \mathcal{M} \Big) \geq 1 - q,\ \forall \mathcal{M}.$$

- Submodel parameter:

$$\beta^*_{\mathcal{M}} = \underset{\alpha, \beta_{\mathcal{M}}}{\arg\min} \sum_{i=1}^{n} \Big(\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\mathcal{M}}^T \beta_{\mathcal{M}} \Big)^2.$$

- Key result (Lee *et al.* 2016): For linear selection rules (e.g. lasso, forward stepwise) $\big(\hat{\beta}_{\hat{\mathcal{M}}}\big)_j \,|\, \mathbf{AY} \leq \mathbf{b}$ is a truncated normal with mean $\big(\beta^*_{\hat{\mathcal{M}}}\big)_j$.
- Relaxation of normality and improvement by randomization (Tian and Taylor 2018).

Background: Post-selection inference

Suppose we have noisy observations of Δ: $Y_i = \Delta(\mathbf{X}_i) + \epsilon_i,\ i = 1, \ldots, n$.

- Model selection procedure: $\{\mathbf{X}_i, Y_i\}_{i=1}^n \mapsto \hat{\mathcal{M}}$.
- **Conditional confidence interval**:

$$\mathrm{P}\Big(\big(\boldsymbol{\beta}^*_{\mathcal{M}}\big)_j \in [D_j^-, D_j^+] \;\Big|\; \hat{\mathcal{M}} = \mathcal{M} \Big) \geq 1 - q,\ \forall \mathcal{M}.$$

- Submodel parameter:

$$\boldsymbol{\beta}^*_{\mathcal{M}} = \underset{\alpha, \boldsymbol{\beta}_{\mathcal{M}}}{\arg\min} \sum_{i=1}^{n} \Big(\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}} \Big)^2.$$

- Key result (Lee *et al.* 2016): For linear selection rules (e.g. lasso, forward stepwise) $\big(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}}\big)_j \,|\, \mathbf{AY} \leq \mathbf{b}$ is a truncated normal with mean $\big(\boldsymbol{\beta}^*_{\hat{\mathcal{M}}}\big)_j$.
- Relaxation of normality and improvement by randomization (Tian and Taylor 2018).

Background: Eliminate the nuisance parameter

- Back to the causal model (of the observables)

$$Y_i = \eta(\boldsymbol{X}_i) + T_i \cdot \Delta(\boldsymbol{X}_i) + \epsilon_i, \;\; i = 1, \ldots, n.$$

- Problem: how to eliminate the nuisance parameter $\eta(\mathbf{x})$?

Robinson (1988)'s transformation

Let $\mu_y(\mathbf{x}) = \mathrm{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}]$ and $\mu_t(\mathbf{x}) = \mathrm{E}[T_i \mid \mathbf{X}_i = \mathbf{x}]$, so $\mu_y(\mathbf{x}) = \eta(\mathbf{x}) + \mu_t(\mathbf{x})\Delta(\mathbf{x})$. An equivalent model is

$$Y_i - \mu_y(\mathbf{X}_i) = \big(T_i - \mu_t(\mathbf{X}_i)\big) \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \;\; i = 1, \ldots, n.$$

- The new nuisance parameters $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ can be directly estimated from the data.
- This is called **R-learning** in an independent work by Nie and Wager (2021).

Background: Eliminate the nuisance parameter

- Back to the causal model (of the observables)

$$Y_i = \eta(\boldsymbol{X}_i) + T_i \cdot \Delta(\boldsymbol{X}_i) + \epsilon_i, \ \ i = 1, \ldots, n.$$

- Problem: how to eliminate the nuisance parameter $\eta(\mathbf{x})$?

Robinson (1988)'s transformation

Let $\mu_y(\mathbf{x}) = \mathrm{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}]$ and $\mu_t(\mathbf{x}) = \mathrm{E}[T_i \mid \mathbf{X}_i = \mathbf{x}]$, so $\mu_y(\mathbf{x}) = \eta(\mathbf{x}) + \mu_t(\mathbf{x})\Delta(\mathbf{x})$. An equivalent model is

$$Y_i - \mu_y(\mathbf{X}_i) = \big(T_i - \mu_t(\mathbf{X}_i)\big) \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \ \ i = 1, \ldots, n.$$

- The new nuisance parameters $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ can be directly estimated from the data.
- This is called **R-learning** in an independent work by Nie and Wager (2021).

Background: Eliminate the nuisance parameter

- Back to the causal model (of the observables)

$$Y_i = \eta(\boldsymbol{X}_i) + T_i \cdot \Delta(\boldsymbol{X}_i) + \epsilon_i, \ \ i = 1, \ldots, n.$$

- Problem: how to eliminate the nuisance parameter $\eta(\mathbf{x})$?

Robinson (1988)'s transformation

Let $\mu_y(\mathbf{x}) = \mathrm{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}]$ and $\mu_t(\mathbf{x}) = \mathrm{E}[T_i \mid \mathbf{X}_i = \mathbf{x}]$, so $\mu_y(\mathbf{x}) = \eta(\mathbf{x}) + \mu_t(\mathbf{x})\Delta(\mathbf{x})$. An equivalent model is

$$Y_i - \mu_y(\mathbf{X}_i) = \big(T_i - \mu_t(\mathbf{X}_i)\big) \cdot \Delta(\mathbf{X}_i) + \epsilon_i, \ \ i = 1, \ldots, n.$$

- The new nuisance parameters $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ can be directly estimated from the data.
- This is called **R-learning** in an independent work by Nie and Wager (2021).

Our complete proposal

1. **Deconfounding:** Estimate $\mu_y(\mathbf{x})$ and $\mu_t(\mathbf{x})$ using machine learning algorithms (e.g. **random forest**).
2. **Model selection:** Select a model for effect modification by solving a **lasso** problem

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \Big[\big(Y_i - \hat{\mu}_y(\mathbf{X}_i)\big) - \big(T_i - \hat{\mu}_t(\mathbf{X}_i)\big) \cdot (\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) \Big]^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

3. **Selective inference:** Use the **pivotal statistic** in Lee *et al.* (2016) to obtain selective confidence intervals of

$$\boldsymbol{\beta}^*_{\hat{\mathcal{M}}} = \underset{\alpha,\boldsymbol{\beta}_{\hat{\mathcal{M}}}}{\arg\min} \sum_{i=1}^{n} \big(T_i - \mu_t(\mathbf{X}_i)\big)^2 \big(\Delta(\mathbf{X}_i) - \alpha - \mathbf{X}_{i,\hat{\mathcal{M}}}^T \boldsymbol{\beta}_{\hat{\mathcal{M}}}\big)^2.$$

Main theoretical result

Assumption

- $\|\hat{\mu}_t - \mu_t\|_2 = o_p(n^{-1/4})$;
- $\|\hat{\mu}_y - \mu_y\|_2 = o_p(1)$;
- $\|\hat{\mu}_t - \mu_t\|_2 \cdot \|\hat{\mu}_y - \mu_y\|_2 = o_p(n^{-1/2})$.

Remark

- Necessary for efficient estimation in partially linear models (Robinson 1988).
- In randomized experiments, $\mu_t(\boldsymbol{x})$ is known.

Theorem

*Under additional assumptions (***boundedness of $|\hat{\mathcal{M}}|$, minimal sparse eigenvalue > 0, smoothness of the pivot***), the selective confidence interval is asymptotically valid.*

Real data example

Motivation: An epidemiological study

- Visser et al. "Elevated C-reactive protein levels in overweight and obese adults". *JAMA* 282, 1999.
- Prespecified subgroup analysis found effect modification by gender. Within women, they found effect modification by age group.

The dataset

- We used a more recent dataset from NHANES 2007–2008 and 2009–2010.
- T: obesity (BMI ≥ 25).
- Y: C-reactive protein level.
- $\mathbf{X}$: gender, age, income, race, marital status, education, vigorous work activity, vigorous recreation activities, smoking, estrogen usage, bronchitis, asthma, emphysema, thyroid, arthritis, heart attack, stroke, liver condition, gout.
- $n = 9677$, $p = 27$ (365 if all the interactions are used).

Results

Our method

1. Use random forest to estimate $\mu_t(\boldsymbol{x})$ and $\mu_y(\boldsymbol{x})$.
2. Use lasso to select a submodel to approximate $\Delta(\boldsymbol{x})$.
3. Use the **truncated normal pivot** to obtain selective CI.

	Estimate	p-value	CI low	CI up	
Gender (Female)	0.476	0.000	0.330	0.624	***
Age	-0.019	0.000	-0.024	-0.015	***
Stroke	-0.515	0.311	-0.899	1.256	
Gout	-0.475	0.493	-0.852	2.295	

(a) Using only main effects to model effect modification.

	Estimate	p-value	CI low	CI up	
Gender (Female)	0.471	0.000	0.323	0.618	***
Age	-0.020	0.000	-0.024	-0.016	***
Age $\times$ Vigorous recreation	0.018	0.371	-0.052	0.027	
Age $\times$ Stroke	-0.036	0.069	-0.054	0.014	.

(b) Using main effects and first-order interactions.

Naive inference is biased

Data snooping

1. Use random forest to estimate $\mu_t(\boldsymbol{x})$ and $\mu_y(\boldsymbol{x})$.
2. Use lasso to select a submodel to approximate $\Delta(\boldsymbol{x})$.
3. Use the **usual regression inference ignoring model selection**:

$$\texttt{lm(Y - } \hat{\mu}_y \sim \texttt{(T - } \hat{\mu}_t\texttt{) X)}.$$

	Estimate	*p*-value	CI low	CI up	
Gender (Female)	0.476	0.000	0.332	0.620	***
Age	-0.019	0.000	-0.023	-0.015	***
Stroke	-0.514	0.016	-0.933	-0.096	*
Gout	-0.473	0.038	-0.919	-0.026	*

(a) Using only main effects to model effect modification.

	Estimate	*p*-value	CI low	CI up	
Gender (Female)	0.471	0.000	0.327	0.615	***
Age	-0.020	0.000	-0.024	-0.016	***
Age $\times$ Vigorous recreation	0.018	0.001	0.008	0.028	***
Age $\times$ Stroke	-0.036	0.000	-0.055	-0.017	***

(b) Using main effects and first-order interactions.

Outline

Post-selection inference for effect modifiers

Sensitivity analysis for individualized treatment rules

Motivation

Background

- Individualized treatment rule (ITR): $r : \mathcal{X} \to \mathcal{A}$.
- **Optimal treatment regime** = ITR with the best *value* $\mathbb{E}[Y(r)]$.
- **Dynamic treatment regimes** = extension to multiple decision points.
- They are central to precision medicine (see e.g. Kosorok and Laber 2019).
- Existing methods usually assume (sequential) unconfoundedness. This allows us to estimate the value of any ITR.

Rosenbaum's sensitivity model

The odds ratio of receiving the treatment for any two individuals with the same observed covariates is bounded between $1/\Gamma$ and Γ (Rosenbaum 1987).

- $\Gamma \geq 1$; $\Gamma = 1$ corresponds to no unmeasured confounders.

Our question

How do we select and rank ITRs under Rosenbaum's sensitivity model?

Motivation

Background

- Individualized treatment rule (ITR): $r : \mathcal{X} \to \mathcal{A}$.
- **Optimal treatment regime** = ITR with the best *value* $\mathbb{E}[Y(r)]$.
- **Dynamic treatment regimes** = extension to multiple decision points.
- They are central to precision medicine (see e.g. Kosorok and Laber 2019).
- Existing methods usually assume (sequential) unconfoundedness. This allows us to estimate the value of any ITR.

Rosenbaum's sensitivity model

The odds ratio of receiving the treatment for any two individuals with the same observed covariates is bounded between $1/\Gamma$ and Γ (Rosenbaum 1987).

- $\Gamma \geq 1$; $\Gamma = 1$ corresponds to no unmeasured confounders.

Our question

How do we select and rank ITRs under Rosenbaum's sensitivity model?

Motivation

Background

- Individualized treatment rule (ITR): $r : \mathcal{X} \to \mathcal{A}$.
- **Optimal treatment regime** = ITR with the best *value* $\mathbb{E}[Y(r)]$.
- **Dynamic treatment regimes** = extension to multiple decision points.
- They are central to precision medicine (see e.g. Kosorok and Laber 2019).
- Existing methods usually assume (sequential) unconfoundedness. This allows us to estimate the value of any ITR.

Rosenbaum's sensitivity model

The odds ratio of receiving the treatment for any two individuals with the same observed covariates is bounded between $1/\Gamma$ and Γ (Rosenbaum 1987).

- $\Gamma \geq 1$; $\Gamma = 1$ corresponds to no unmeasured confounders.

Our question

How do we select and rank ITRs under Rosenbaum's sensitivity model?

Key conclusion: Value $\neq$ Robustness

The estimated value from some observational data assuming ignorability is a poor indicator for robustness.

A counter-intuitive example

Let $r_2 \succ_\Gamma r_1$ or simply $r_2 \succ r_1$ denote that the value of r_2 is *always* greater than r_1 under the Γ-sensitivity model.
Then, it is possible that

- Under $\Gamma = 1$, $r_2 \succ r_1 \succ r_0$ (so $r_2 \succ r_0$);
- Under some $\Gamma > 1$, $r_1 \succ r_0$ but $r_2 \not\succ r_0$.

Why?

- Value is only **partially identified** in Rosenbaum's (and other) sensitivity model.
- So value only induces a **partial order** between ITRs.

Key conclusion: Value $\neq$ Robustness

The estimated value from some observational data assuming ignorability is a poor indicator for robustness.

A counter-intuitive example

Let $r_2 \succ_\Gamma r_1$ or simply $r_2 \succ r_1$ denote that the value of r_2 is *always* greater than r_1 under the Γ-sensitivity model.
Then, it is possible that

- Under $\Gamma = 1$, $r_2 \succ r_1 \succ r_0$ (so $r_2 \succ r_0$);
- Under some $\Gamma > 1$, $r_1 \succ r_0$ but $r_2 \not\succ r_0$.

Why?

- Value is only **partially identified** in Rosenbaum's (and other) sensitivity model.
- So value only induces a **partial order** between ITRs.

Key conclusion: Value $\neq$ Robustness

The estimated value from some observational data assuming ignorability is a poor indicator for robustness.

A counter-intuitive example

Let $r_2 \succ_\Gamma r_1$ or simply $r_2 \succ r_1$ denote that the value of r_2 is *always* greater than r_1 under the Γ-sensitivity model.
Then, it is possible that

- Under $\Gamma = 1$, $r_2 \succ r_1 \succ r_0$ (so $r_2 \succ r_0$);
- Under some $\Gamma > 1$, $r_1 \succ r_0$ but $r_2 \not\succ r_0$.

Why?

- Value is only **partially identified** in Rosenbaum's (and other) sensitivity model.
- So value only induces a **partial order** between ITRs.

Notation

Running example: Malaria in West Africa

Dataset from Hsu *et al.* (2013): 1560 matched pairs of Nigerians.

- Treatment $A \in \mathcal{A} = \{0, 1\}$. $A = 1$: receives treatment (insecticide spray + drug).
- Covariates $X \in \mathcal{X}$ (gender and age);
- Outcome Y (amount of malaria-causing parasites in blood).
- ITR $r : \mathcal{X} \to \mathcal{A}$ (six rules: $r_0, r_1, \ldots, r_5$, where r_i assigns treatment to the youngest $i \times 20\%$.)
- Potential outcomes $Y(0)$ and $Y(1)$, so $Y(r) = Y(0)1_{\{r(X)=0\}} + Y(1)1_{\{r(X)=1\}}$.
- **Value function** $V(r) = \mathbb{E}[Y(r)]$.

Comparing two ITRs

No unmeasured confoudners

- The **value difference** is $V(r_2) - V(r_1) = \mathbb{E}[Y(r_2) - Y(r_1) \,|\, r_2 \neq r_1] \cdot \mathbb{P}(r_2 \neq r_1)$.
- In our example (nested ITRs),
  $V(r_2) - V(r_1) = \mathbb{E}[Y(1) - Y(0) \,|\, \text{Age} \in [7, 20)] \cdot \mathbb{P}(\text{Age} \in [7, 20))$.
- Point identified under standard assumptions (consistency, unconfoundedness, positivity).

Unmeasured confoudners

- Define $r_1 \prec_{\Gamma,\delta} r_2$ if $V(r_2) - V(r_1) > \delta$ **for all distributions in the Γ-sensitivity model**.
  Can verify this is a **partial order**.
- Can be tested adapting the studentized test for Neyman's weak null (Fogarty 2020).

Comparing two ITRs

No unmeasured confoudners

- The **value difference** is $V(r_2) - V(r_1) = \mathbb{E}[Y(r_2) - Y(r_1) \,|\, r_2 \neq r_1] \cdot \mathbb{P}(r_2 \neq r_1)$.
- In our example (nested ITRs),
  $V(r_2) - V(r_1) = \mathbb{E}[Y(1) - Y(0) \,|\, \text{Age} \in [7, 20)] \cdot \mathbb{P}(\text{Age} \in [7, 20))$.
- Point identified under standard assumptions (consistency, unconfoundedness, positivity).

Unmeasured confoudners

- Define $r_1 \prec_{\Gamma,\delta} r_2$ if $V(r_2) - V(r_1) > \delta$ **for all distributions in the Γ-sensitivity model**.
  Can verify this is a **partial order**.
- Can be tested adapting the studentized test for Neyman's weak null (Fogarty 2020).

Comparing multiple ITRs

Related problem: selecting subpopulations

- Suppose we observe $Y_i \overset{\text{ind.}}{\sim} \mathrm{N}(\mu_i, 1)$ for subpopulation i.
- Gibbons *et al.* (1999) has defined seven possible goals for ranking and selecting subpopulations.

Our problem

Given $\mathcal{R} = \{r_0, r_1, \ldots, r_K\}$, three goals are relevant for comparing multiple ITRs:

1. What is the ordering of all the ITRs?
2. Which ITRs are among the best?
3. Which ITRs are better than the control rule r_0?

Cannot directly use existing methods because $\prec_\Gamma$ is only a partial order.

Comparing multiple ITRs

Related problem: selecting subpopulations

- Suppose we observe $Y_i \overset{\text{ind.}}{\sim} \mathrm{N}(\mu_i, 1)$ for subpopulation i.
- Gibbons *et al.* (1999) has defined seven possible goals for ranking and selecting subpopulations.

Our problem

Given $\mathcal{R} = \{r_0, r_1, \ldots, r_K\}$, three goals are relevant for comparing multiple ITRs:

1. What is the ordering of all the ITRs?
2. Which ITRs are among the best?
3. Which ITRs are better than the control rule r_0?

Cannot directly use existing methods because $\prec_\Gamma$ is only a partial order.

Comparing multiple ITRs

Some definitions

- The **maximal rules** $\mathcal{R}_{\max,\Gamma}$ are the ones not dominated by others.
- The **positive rules** $\mathcal{R}_{\text{pos},\Gamma}$ (or null rules $\mathcal{R}_{\text{nul},\Gamma}$) are the ones which dominate (or don't dominate) the control rule r_0.

Possible objectives

1. Construct a set of ordered ITR pairs, $\hat{\mathcal{O}}_\Gamma \subset \{(r_i, r_j),\ i, j = 0, \ldots, K,\ i \neq j\}$, such that

$$\mathbb{P}(r_i \prec_\Gamma r_j,\ \forall (r_i, r_j) \in \hat{\mathcal{O}}_\Gamma) \geq 1 - \alpha.$$

2. Construct $\hat{\mathcal{R}}_{\max,\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \geq 1 - \alpha$.
3. Construct $\hat{\mathcal{R}}_{\text{pos},\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\hat{\mathcal{R}}_{\text{pos},\Gamma} \cap \mathcal{R}_{\text{null},\Gamma} = \emptyset) \geq 1 - \alpha$.

Proposed solution

Use multiple testing procedures that control the family-wise error rate, but use a planning sample to reduce the number of tests (Heller *et al.* 2009; Zhao *et al.* 2018).

Comparing multiple ITRs

Some definitions

- The **maximal rules** $\mathcal{R}_{\max,\Gamma}$ are the ones not dominated by others.
- The **positive rules** $\mathcal{R}_{\text{pos},\Gamma}$ (or null rules $\mathcal{R}_{\text{nul},\Gamma}$) are the ones which dominate (or don't dominate) the control rule r_0.

Possible objectives

1. Construct a set of ordered ITR pairs, $\hat{\mathcal{O}}_\Gamma \subset \{(r_i, r_j),\ i, j = 0, \ldots, K,\ i \neq j\}$, such that

$$\mathbb{P}(r_i \prec_\Gamma r_j,\ \forall (r_i, r_j) \in \hat{\mathcal{O}}_\Gamma) \geq 1 - \alpha.$$

2. Construct $\hat{\mathcal{R}}_{\max,\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \geq 1 - \alpha$.
3. Construct $\hat{\mathcal{R}}_{\text{pos},\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\hat{\mathcal{R}}_{\text{pos},\Gamma} \cap \mathcal{R}_{\text{null},\Gamma} = \emptyset) \geq 1 - \alpha$.

Proposed solution

Use multiple testing procedures that control the family-wise error rate, but use a planning sample to reduce the number of tests (Heller *et al.* 2009; Zhao *et al.* 2018).

Comparing multiple ITRs

Some definitions

- The **maximal rules** $\mathcal{R}_{\max,\Gamma}$ are the ones not dominated by others.
- The **positive rules** $\mathcal{R}_{\text{pos},\Gamma}$ (or null rules $\mathcal{R}_{\text{nul},\Gamma}$) are the ones which dominate (or don't dominate) the control rule r_0.

Possible objectives

1. Construct a set of ordered ITR pairs, $\hat{\mathcal{O}}_\Gamma \subset \{(r_i, r_j),\ i, j = 0, \ldots, K,\ i \neq j\}$, such that

$$\mathbb{P}(r_i \prec_\Gamma r_j,\ \forall (r_i, r_j) \in \hat{\mathcal{O}}_\Gamma) \geq 1 - \alpha.$$

2. Construct $\hat{\mathcal{R}}_{\max,\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \geq 1 - \alpha$.
3. Construct $\hat{\mathcal{R}}_{\text{pos},\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\hat{\mathcal{R}}_{\text{pos},\Gamma} \cap \mathcal{R}_{\text{null},\Gamma} = \emptyset) \geq 1 - \alpha$.

Proposed solution

Use multiple testing procedures that control the family-wise error rate, but use a planning sample to reduce the number of tests (Heller *et al.* 2009; Zhao *et al.* 2018).

Objective 1: Ordered pairs

Malaria example: denote $H_{ij}: \ r_i \not\prec_\Gamma r_j$.

Ordered hypotheses after using the planning sample

- $\Gamma = 1$: $H_{01}, H_{02}, H_{03}, H_{04}, H_{05}, H_{13}, H_{12}, H_{14}, H_{15}, H_{23}, \ldots$.
- $\Gamma = 2$: $H_{02}, H_{01}, H_{03}, H_{04}, H_{05}, H_{12}, H_{13}, H_{14}, H_{15}, H_{45}, \ldots$.

$|\hat{\mathcal{O}}| = 5$ $|\hat{\mathcal{O}}| = 7$

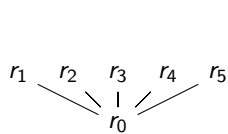
Hasse diagrams for $\Gamma = 2$: Bonferroni's correction (left) and our proposal (right).

Objective 1: Ordered pairs

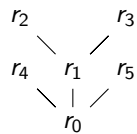
Malaria example: denote $H_{ij}: \ r_i \not\preceq_\Gamma r_j$.

Ordered hypotheses after using the planning sample

- $\Gamma = 1$: $H_{01}, H_{02}, H_{03}, H_{04}, H_{05}, H_{13}, H_{12}, H_{14}, H_{15}, H_{23}, \ldots$.
- $\Gamma = 2$: $H_{02}, H_{01}, H_{03}, H_{04}, H_{05}, H_{12}, H_{13}, H_{14}, H_{15}, H_{45}, \ldots$.



$|\hat{\mathcal{O}}| = 5$



$|\hat{\mathcal{O}}| = 7$

Hasse diagrams for $\Gamma = 2$: Bonferroni's correction (left) and our proposal (right).

Objective 2 & 3

Selecting maximal ITRs

- Key observation: $\mathbb{P}(r_i \not\succ_\Gamma r_j \text{ is rejected} \mid r_i \in \mathcal{R}_{\max,\Gamma}) \leq \alpha$.
- This motivates us to use all the "leaves" in the Hasse diagram as the maximal elements.

- This satisfies $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \not\subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \leq \alpha$ if the FWER for $\hat{\mathcal{O}}_\Gamma$ is less than α.

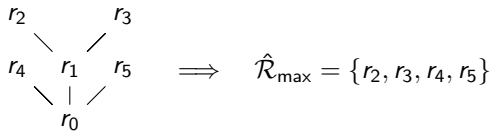
Selecting positive ITRs

- Simply needs to test the hypotheses $H_{0i} : r_0 \not\succ_\Gamma r_i,\ i = 1, \ldots, K$.
- Use the same multiple testing procedure as before.

Objective 2 & 3

Selecting maximal ITRs

- Key observation: $\mathbb{P}(r_i \not\succ_\Gamma r_j \text{ is rejected} \mid r_i \in \mathcal{R}_{\max,\Gamma}) \leq \alpha$.
- This motivates us to use all the "leaves" in the Hasse diagram as the maximal elements.



- This satisfies $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \not\subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \leq \alpha$ if the FWER for $\hat{\mathcal{O}}_\Gamma$ is less than α.

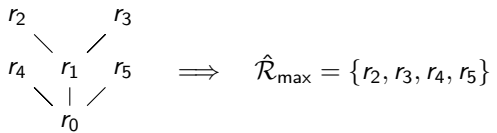
Selecting positive ITRs

- Simply needs to test the hypotheses $H_{0i} : r_0 \not\succ_\Gamma r_i,\ i = 1, \ldots, K$.
- Use the same multiple testing procedure as before.

Objective 2 & 3

Selecting maximal ITRs

- Key observation: $\mathbb{P}(r_i \not\succ_\Gamma r_j \text{ is rejected} \,|\, r_i \in \mathcal{R}_{\max,\Gamma}) \leq \alpha$.
- This motivates us to use all the "leaves" in the Hasse diagram as the maximal elements.



- This satisfies $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \not\subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \leq \alpha$ if the FWER for $\hat{\mathcal{O}}_\Gamma$ is less than α.

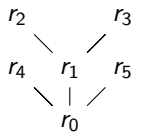
Selecting positive ITRs

- Simply needs to test the hypotheses $H_{0i} : r_0 \not\succ_\Gamma r_i,\ i = 1, \ldots, K$.
- Use the same multiple testing procedure as before.

Objective 2 & 3

Selecting maximal ITRs

- Key observation: $\mathbb{P}(r_i \not\succ_\Gamma r_j \text{ is rejected} \mid r_i \in \mathcal{R}_{\max,\Gamma}) \leq \alpha$.
- This motivates us to use all the "leaves" in the Hasse diagram as the maximal elements.



$$\implies \quad \hat{\mathcal{R}}_{\max} = \{r_2, r_3, r_4, r_5\}$$

- This satisfies $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \not\subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \leq \alpha$ if the FWER for $\hat{\mathcal{O}}_\Gamma$ is less than α.

Selecting positive ITRs

- Simply needs to test the hypotheses $H_{0i} : r_0 \not\succ_\Gamma r_i,\ i = 1, \ldots, K$.
- Use the same multiple testing procedure as before.

Malaria example: Results

Γ	$\hat{\mathcal{R}}_{\max,\Gamma}$	$\hat{\mathcal{R}}_{\text{pos},\Gamma}$
1.0	$\{r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$
1.5	$\{r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$
3.5	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3\}$
4.0	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2\}$
6.0	$\{r_0, r_1, r_2, r_3, r_4, r_5\}$	$\emptyset$

- A more complicated example can be found in the paper.

Discussion

- Another consideration in decision making: **Reliability/robustness of causal inference and individulized treatment**.
  - Post-selection inference for effect modifiers;
  - Selecting ITRs with unmeasured confounders.
- This talk only considered the most classical settings (binary treatment, linear model, Rosenbaum's sensitivity model).
- Selective inference for partially identified/ordered problems: a potentially new topic?

References

1. S. Athey, J. Tibshirani, S. Wager, *Annals of Statistics* **47**, 1148–1178 (2 2019).
2. C. B. Fogarty, *Journal of the American Statistical Association* **115**, 1518–1530 (531 2020).
3. J. D. Gibbons, I. Olkin, M. Sobel, *Selecting and ordering populations: A new statistical methodology*, (SIAM, 2nd, 1999).
4. D. P. Green, H. L. Kern, presented at The annual summer meeting of the society of political methodology.
5. D. E. Grobbee, A. W. Hoes, *Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research*, (Jones & Bartlett Learning, 2009).
6. R. Heller, P. R. Rosenbaum, D. S. Small, *Journal of the American Statistical Association* **104**, 1090–1101 (2009).
7. J. L. Hill, *Journal of Computational and Graphical Statistics* **20**, 217–240 (2011).
8. J. Y. Hsu, D. S. Small, P. R. Rosenbaum, *Journal of the American Statistical Association* **108**, 135–148 (2013).
9. J. Y. Hsu, J. R. Zubizarreta, D. S. Small, P. R. Rosenbaum, *Biometrika* **102**, 767–782 (2015).
10. K. Imai, M. Ratkovic, *The Annals of Applied Statistics* **7**, 443–470 (2013).
11. M. R. Kosorok, E. B. Laber, *Annual Review of Statistics and Its Application* **6**, 263–286 (2019).
12. J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, *Annals of Statistics* **44**, 907–927 (2016).
13. S. A. Murphy, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 331–355 (2003).
14. X. Nie, S. Wager, *Biometrika* **108**, 299–319 (2021).
15. S. Powers *et al.*, *Statistics in medicine* **37**, 1767–1787 (2018).
16. M. Qian, S. A. Murphy, *Annals of statistics* **39**, 1180 (2011).
17. P. M. Robinson, *Econometrica* **56**, 931–954 (1988).
18. P. R. Rosenbaum, *Biometrika* **74**, 13–26 (1987).
19. E. A. Stuart, S. R. Cole, C. P. Bradshaw, P. J. Leaf, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**, 369–386 (2011).
20. L. Tian, A. Alizadeh, A. Gentles, R. Tibshirani, *Journal of the American Statistical Association* **109**, 1517–1532 (2014).
21. X. Tian, J. Taylor, *Annals of Statistics* **46**, 679–710 (2018).
22. T. J. VanderWeele, J. M. Robins, *Epidemiology* **18**, 561–568 (2007).
23. Q. Zhao, D. S. Small, P. R. Rosenbaum, *Journal of the American Statistical Association* **113**, 1070–1084 (2018).
24. Y. Zhao, D. Zeng, A. J. Rush, M. R. Kosorok, *Journal of the American Statistical Association* **107**, 1106–1118 (2012).