

Lecture 5: Heterogeneous Treatment Effects: From One Number to a Function

DATA 8020: Advanced Causal Inference

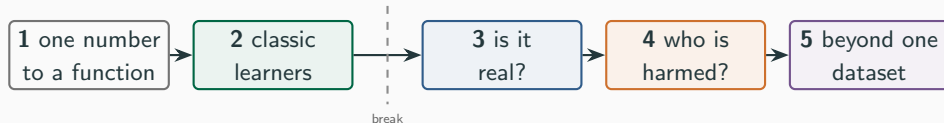
Jin-Hong Du

Institute of Data Science, The University of Hong Kong

October 2, 2026

Today: who does the treatment help, and how would we know?

Five parts: classic methods before the break, recent directions after it.



By the end, you should be able to:

- say what a CATE is, and why it is not one person's effect;
- describe the classic learners and the first way each fails;
- check whether estimated heterogeneity is real before reporting it; and
- name three open directions and the question each answers.

A varying estimate is not evidence of a varying effect.

Lecture 4 gave one number; today it becomes a function

	Lecture 4	Lecture 5
Question	how large is the effect?	for whom is it large?
Target	one number, τ	a function, $\tau(x)$
Tools	AIPW score, cross-fitting	the same score and folds

$$\tau = \mathbb{E}\{Y(1) - Y(0)\} \text{ becomes } \tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\}.$$

Colours carry the same roles as before: green treatment side, blue control side, orange what the data cannot show you, purple the estimand and estimator.

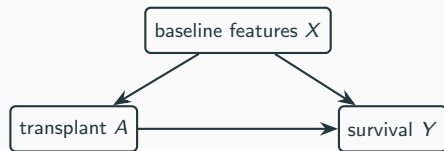
One number to a function

One average can describe nobody in the room

Group	Share	True effect	What the ATE says
Least severe quartile	25%	−0.41 years	+0.97 years
Middle two quartiles	50%	+0.98 years	
Most severe quartile	25%	+2.32 years	

Subgroup / individual effects are heterogeneous.

The estimand: a conditional average treatment effect (CATE)



Same three assumptions as Lecture 4:

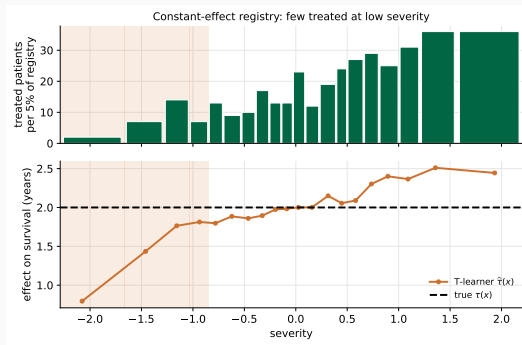
- consistency;
- no unmeasured confounding, $Y(a) \perp\!\!\!\perp A \mid X$;
- overlap, at every x as before.

$$\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\} = \mu_1(x) - \mu_0(x), \quad \mu_a(x) = \mathbb{E}[Y \mid A = a, X = x].$$

Also called the **heterogeneous treatment effect**.

Each x is now its own estimand.

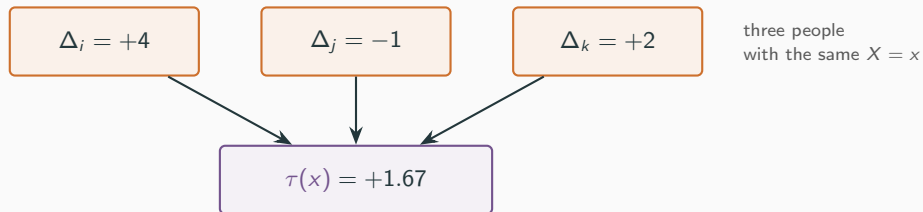
Where overlap fails, the curve is pure extrapolation



Shaded: fewer than one patient in eight treated. The curve still prints a number there, from extrapolation alone.

Trimming the ATE to an overlap region S already gave a CATE, $\mathbb{E}\{Y(1) - Y(0) \mid X \in S\}$.
Today S shrinks to $\{x\}$.

A CATE averages people; it is not a person



One person's effect $\Delta_i = Y_i(1) - Y_i(0)$ is never observed: we see one of the two terms, at most.

Conditioning on more X narrows the **group**. It never turns the average into a person.

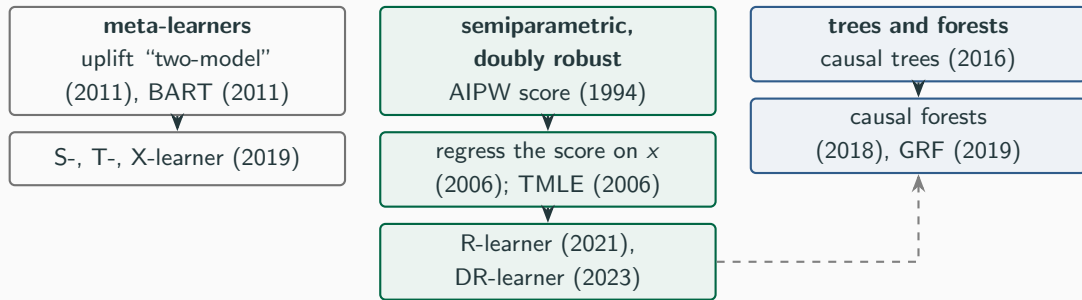
The belief this lecture tests

*"My CATE model fits well and its estimates vary across patients.
So the effect varies, and the top group benefits most."*



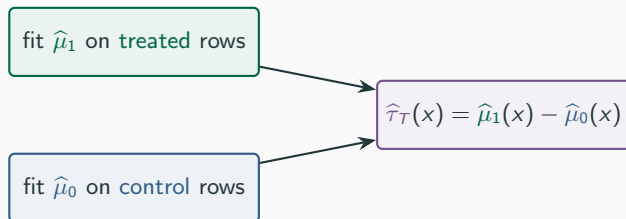
Classic learners

Where the names come from: three lines, now merging



- **Letters are recent labels for older ideas.** S, T, X distinguish nuisance models; R is Robinson's partialling-out (Robinson, 1988), the CATE version of DML; DR regresses Lecture 4's score on x .
- **TMLE is a sibling of the DR-learner, not its source.** Both descend from the Robins–van der Laan line; TMLE targets one number. GRF now centres with the R-learner, and orthogonal learning unifies R and DR (Foster and Syrgkanis, 2023; Semenova and Chernozhukov, 2021).

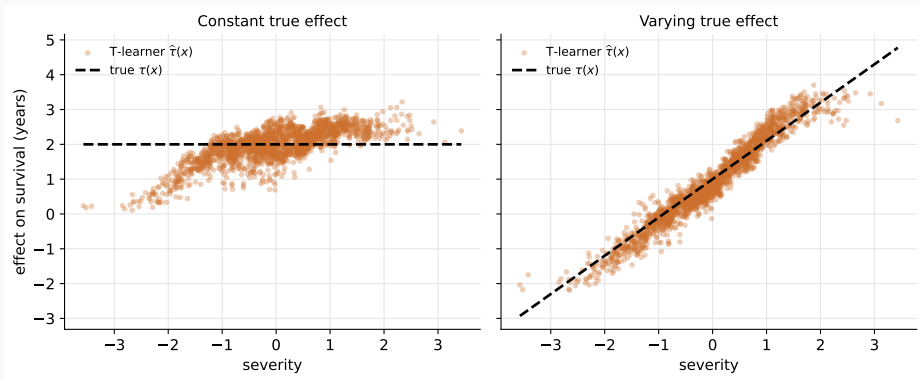
The obvious estimator: fit both arms and subtract



This is the **T-learner**, T for *two* models. The **S-learner**, S for a *single* model, fits one model on (X, A) and flips A .

Both work if both fits are good. At realistic sample sizes they often are not.

Two good fits, one bad difference



Left: the true effect is 2.00 years for everyone, yet the T-learner's estimate slopes (SD 0.48 years). Right: the effect truly varies.

Few treated rows, so $\hat{\mu}_1$ is smoother than $\hat{\mu}_0$; their difference inherits that gap (Künzel et al., 2019).

X-learner: three steps that borrow the bigger arm

1. **Fit both arms** as in the T-learner: $\hat{\mu}_1$ on **treated** rows, $\hat{\mu}_0$ on **control** rows.
2. **Cross over** (the X): impute each unit's effect with the **other** arm's fit, then regress it on X within each arm.

$$\tilde{\Delta}_i = \begin{cases} Y_i - \hat{\mu}_0(X_i), & A = 1 \Rightarrow \text{regress on } X_i \text{ to get } \hat{\tau}_1(x), \\ \hat{\mu}_1(X_i) - Y_i, & A = 0 \Rightarrow \text{regress on } X_i \text{ to get } \hat{\tau}_0(x). \end{cases}$$

3. **Blend** with the propensity, weighting toward the arm-specific estimate built on the larger arm:

$$\hat{\tau}_X(x) = \hat{e}(x) \hat{\tau}_0(x) + \{1 - \hat{e}(x)\} \hat{\tau}_1(x).$$

Small treated arm $\Rightarrow \hat{e}(x)$ small $\Rightarrow$ the blend leans on $\hat{\tau}_1$, which uses treated outcomes against $\hat{\mu}_0$ from the big control arm. The poorly fitted $\hat{\mu}_1$ barely enters.

No sample splitting is needed for step 2: each unit is imputed by the *other* arm's fit, so its own outcome never trained the model applied to it. Cross-fit anyway, so that $\hat{e}$ and $\hat{\tau}_X$ are held out for Part 3.

Lecture 4's score already contains $\tau(x)$

The AIPW score, unchanged:

$$\phi(O) = \mu_1(X) - \mu_0(X) + \frac{A}{e(X)} \{Y - \mu_1(X)\} - \frac{1-A}{1-e(X)} \{Y - \mu_0(X)\}.$$

average it: $\mathbb{E}\{\phi(O)\} = \tau$ (Lecture 4)

regress it on X : $\mathbb{E}\{\phi(O) \mid X = x\} = \tau(x)$ (**DR-learner**)

The score divides by $e(x)$: where overlap is thin, the regression is noisy (Kennedy, 2023).

R-learner: residual on residual

Take the marginal fit $m(x) = \mathbb{E}[Y \mid X = x]$ and the propensity $e(x)$. Then

$$Y - m(X) = \{A - e(X)\} \tau(X) + \text{noise}.$$

So choose $\hat{\tau}$ to make that fit as good as possible, with cross-fitted $\hat{m}$ and $\hat{e}$ (Robinson, 1988; Nie and Wager, 2021).

R-learner: residual on residual

Take the marginal fit $m(x) = \mathbb{E}[Y \mid X = x]$ and the propensity $e(x)$. Then

$$Y - m(X) = \{A - e(X)\} \tau(X) + \text{noise}.$$

So choose $\hat{\tau}$ to make that fit as good as possible, with cross-fitted $\hat{m}$ and $\hat{e}$ (Robinson, 1988; Nie and Wager, 2021).

- **Binary A :** no model for Y is assumed. $\mathbb{E}[Y \mid A, X] = \mu_0(X) + A \tau(X)$ holds for any binary regressor, so the identity above is exact.
- **General A :** the identity needs the effect linear in A , $\mathbb{E}[Y \mid A, X] = g(X) + A \tau(X)$, a partially linear model with a varying coefficient. Robinson's model is the special case $\tau(X) \equiv \tau$.

R-learner: residual on residual

Take the marginal fit $m(x) = \mathbb{E}[Y \mid X = x]$ and the propensity $e(x)$. Then

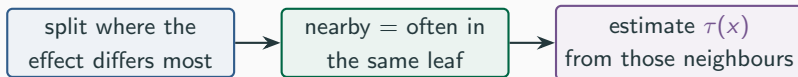
$$Y - m(X) = \{A - e(X)\} \tau(X) + \text{noise}.$$

So choose $\hat{\tau}$ to make that fit as good as possible, with cross-fitted $\hat{m}$ and $\hat{e}$ (Robinson, 1988; Nie and Wager, 2021).

- **Binary** A : no model for Y is assumed. $\mathbb{E}[Y \mid A, X] = \mu_0(X) + A \tau(X)$ holds for any binary regressor, so the identity above is exact.
- **General** A : the identity needs the effect linear in A , $\mathbb{E}[Y \mid A, X] = g(X) + A \tau(X)$, a partially linear model with a varying coefficient. Robinson's model is the special case $\tau(X) \equiv \tau$.

Same condition as Lecture 4: if $\hat{m}$ and $\hat{e}$ are good enough, it does as well as if we knew them.

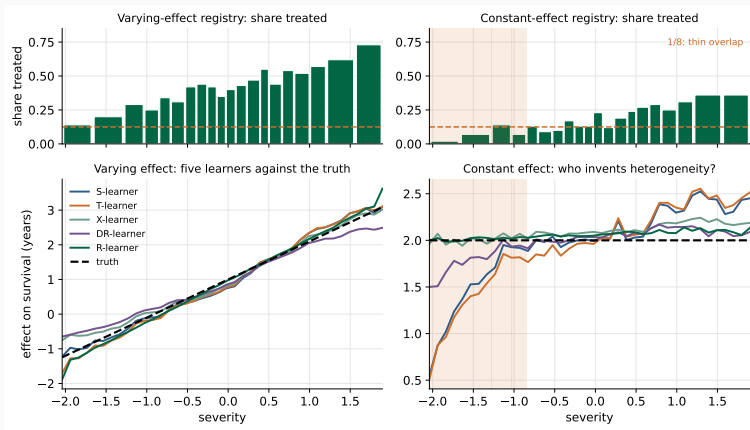
Causal forests: let the data choose the neighbourhood



Honest: the rows that choose a split do not estimate inside it — the cross-fitting rule of Lecture 4, applied to leaves (Wager and Athey, 2018; Athey et al., 2019).

The learner built to give a confidence interval at each x .

Five learners, two registries



Top: share treated per bin; each bin holds 5% of patients, so widths vary; shaded below one in eight. Bottom: binned averages, **different vertical scales**. Learners disagree where overlap is thin: the DR-learner divides by $\hat{e}$ there, the S- and T-learner extrapolate. The R-learner's linear working model matches the truth. **Read the mechanism, not a ranking.**

Which learner, when

	want a point estimate a ranking needs only this	want an interval at each x
arms balanced	DR- or R-learner	causal forest, or DR/R with an inferential 2nd stage
arms unbalanced	X-learner	same, check overlap first

Choosing a learner is choosing an assumption. None of these is a default.

Short break

After the break: recent directions —
is it real, who is harmed, and what if we have two datasets?

Is it real?

Validation without ground truth: one trick, three questions

The truth $\tau(x)$ is never observed, so no error can be computed. But on **held-out** rows the cross-fitted score is an unbiased stand-in:

$\mathbb{E}[\phi_i | X_i] = \tau(X_i)$ so every check below compares held-out $\hat{\phi}_i$ with something built from $\hat{\tau}$.

	Question	Used when you will. . .
Q1 calibration	do the values of $\hat{\tau}$ track real variation?	report $\tau(x)$ itself
Q2 ranking	is the order right, even if the values are not?	act on a ranking (Lecture 6)
Q3 model selection	which learner to trust, with no error to compute?	choose the $\hat{\tau}$ to carry forward

Pitfall on the way: a subgroup chosen because it looked best in-sample but looks worse on new test data. That is why “held-out” is in every line.

A check that does not need the truth

Q1 (calibration): do the values of $\hat{\tau}$ track real variation?

Fit $\hat{\tau}(x)$ on one half. On the **other** half, with cross-fitted scores $\hat{\phi}_i$, $\mathbb{E}[\phi_i | X_i] = \tau(X_i)$, run

$$\hat{\phi}_i = \alpha \cdot \bar{\tau} + \beta \cdot \{\hat{\tau}(X_i) - \bar{\tau}\} + \varepsilon_i, \quad \beta = \frac{\text{Cov}\{\tau(X), \hat{\tau}(X)\}}{\text{Var}\{\hat{\tau}(X)\}}.$$

A check that does not need the truth

Q1 (calibration): do the values of $\hat{\tau}$ track real variation?

Fit $\hat{\tau}(x)$ on one half. On the **other** half, with cross-fitted scores $\hat{\phi}_i$, $\mathbb{E}[\phi_i | X_i] = \tau(X_i)$, run

$$\hat{\phi}_i = \alpha \cdot \bar{\hat{\tau}} + \beta \cdot \{\hat{\tau}(X_i) - \bar{\hat{\tau}}\} + \varepsilon_i, \quad \beta = \frac{\text{Cov}\{\tau(X), \hat{\tau}(X)\}}{\text{Var}\{\hat{\tau}(X)\}}.$$

Test. $H_0: \beta = 0$ vs. $H_1: \beta > 0$.

$t = \hat{\beta} / \widehat{\text{SE}}(\hat{\beta})$ with robust SE;

reject at the 5% level if $t > 1.64$.

Read. $\alpha \approx 1$: the **average** is right.

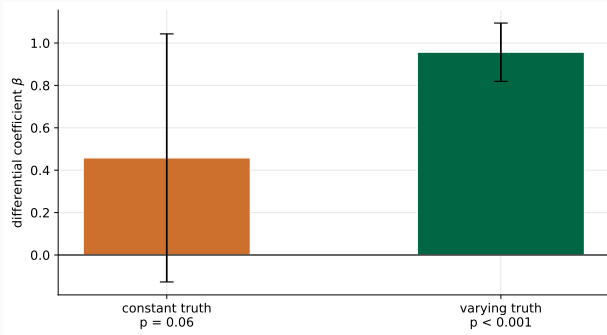
$\beta > 0$: the **variation** is real;

CI for β covers 1: well scaled; $\hat{\beta} < 1$: shrink.

$\hat{\tau}$ is fixed by the training half, so this is ordinary regression inference on the test half.

Formally: the *best linear predictor* (BLP) of $\tau(X)$ given $\hat{\tau}(X)$, Chernozhukov et al. (2025); in `grf`, `test_calibration`.

The check, on both registries



- Constant truth: $\hat{\beta} = 0.46$, $p = 0.06$ — a **near miss**, not a clean pass.
- Varying truth: $\hat{\beta} = 0.96$, $p < 0.001$. In both, $\hat{\alpha} = 0.96$.
- Use robust standard errors: the score is heavy-tailed where $\hat{e}$ is small.

DR-learner, split sample. Failing to reject means “not detected with this X and this n ”, never “everyone is alike”.

Ranking: does treating the top first pay off?

Q2 (ranking): is the *order* right, even if the values are not?

On the **held-out** half, sort by $\hat{\tau}(x)$, treat the top fraction q , and compare with treating at random:

$$\text{TOC}(q) = \mathbb{E}\{\tau(X) \mid \hat{\tau}(X) \text{ in the top } q\} - \mathbb{E}\{\tau(X)\}, \quad \text{RATE} = \int_0^1 \text{TOC}(q) dq.$$

Test. $H_0: \text{RATE} = 0$ vs. $H_1: \text{RATE} > 0$.

$t = \widehat{\text{RATE}} / \widehat{\text{SE}}$, bootstrap SE over rows;

reject at the 5% level if $t > 1.64$.

Read. $\text{TOC}(q) > 0$: top q beats random;

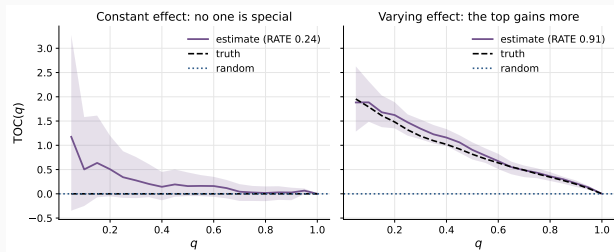
$\text{TOC}(1) = 0$ by construction;

$\text{RATE} > 0$: the order finds someone special.

Estimated with $\hat{\phi}_i$ in place of $\tau(X_i)$. Only the order enters, so Q2 can pass where Q1 fails.

TOC = targeting operator characteristic; RATE = rank-weighted average treatment effect (Yadlowsky et al., 2025). The Qini curve of uplift modelling is the same idea, weighted by q .

The ranking, on both registries

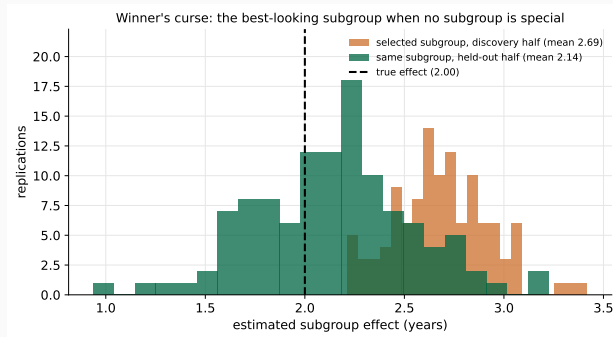


- Constant truth: RATE 0.24 (SE 0.16), $p = 0.07$, where the truth is 0 — the same **near miss** as Q1.
- Varying truth: RATE 0.91 (0.07); the same ranking truly delivers 0.85.
- Band: bootstrap 95%. Dashed: what this ranking truly delivers.

DR-learner ranking, held-out half. A ranking is all a budgeted policy uses (Lecture 6).

The best-looking subgroup is usually lucky

Pitfall (winner's curse): is the subgroup I found real, or just selected?

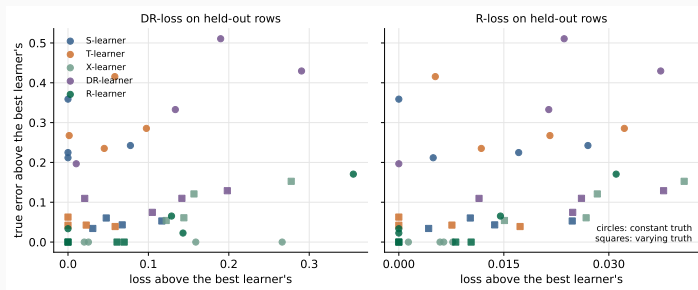


- True effect is constant 2.00. The best of ten looks like 2.69; held out, it is 2.14.
- Report the **held-out** number for any subgroup chosen from the data.

114 replications of the constant-effect registry.

Choosing a learner without the truth

Q3 (model selection): which learner should I trust, when I cannot compute its error?



- **Core idea:** score each learner on held-out rows with a loss built from $\hat{\phi}_i$ or from residuals.
- Here each loss picked the truly best learner in 2 of 8 samples.

Open problem: no criterion wins everywhere.

Formally: *proxy losses* for CATE model selection, Curth and van der Schaar (2023). Use them to rule learners out, not to crown one.

Who is harmed?

The tempting next step: label individuals by $\hat{\tau}(x)$

$\hat{\tau}$ passed Part 3. Can we now say *who* is harmed, and *what* each patient gains? Three natural choices:

The tempting next step: label individuals by $\hat{\tau}(x)$

$\hat{\tau}$ passed Part 3. Can we now say *who* is harmed, and *what* each patient gains? Three natural choices:

1. **Count the harmed:** the proportion of patients with $\hat{\tau}(x) < 0$.
2. **Pick whom to treat:** everyone with $\hat{\tau}(x) > c$.
3. **Tell a patient:** “your effect is $\hat{\tau}(x)$.”

The tempting next step: label individuals by $\hat{\tau}(x)$

$\hat{\tau}$ passed Part 3. Can we now say *who* is harmed, and *what* each patient gains? Three natural choices:

1. **Count the harmed:** the proportion of patients with $\hat{\tau}(x) < 0$.
2. **Pick whom to treat:** everyone with $\hat{\tau}(x) > c$.
3. **Tell a patient:** “your effect is $\hat{\tau}(x)$.”

People differ within x : $\Delta_i = Y_i(1) - Y_i(0) = \tau(X_i) + U_i$, $\mathbb{E}[U_i | X_i] = 0$, U_i never observed.

The tempting next step: label individuals by $\hat{\tau}(x)$

$\hat{\tau}$ passed Part 3. Can we now say *who* is harmed, and *what* each patient gains? Three natural choices:

1. **Count the harmed:** the proportion of patients with $\hat{\tau}(x) < 0$.
2. **Pick whom to treat:** everyone with $\hat{\tau}(x) > c$.
3. **Tell a patient:** “your effect is $\hat{\tau}(x)$.”

People differ within x : $\Delta_i = Y_i(1) - Y_i(0) = \tau(X_i) + U_i$, $\mathbb{E}[U_i | X_i] = 0$, U_i never observed.

- Choice 2 survives: a rule is judged by its *mean* gain, $\mathbb{E}[\Delta_i \mathbf{1}\{\tau(X_i) > c\}] = \mathbb{E}[\tau(X) \mathbf{1}\{\tau(X) > c\}]$, and U_i averages out (Lecture 6).
- Choices 1 and 3 need the spread of U_i within x , e.g. $\mathbb{P}(\Delta_i < 0 | x) = \mathbb{P}\{U_i < -\tau(x) | x\}$. We see one of $Y_i(1)$, $Y_i(0)$, never both: that spread is **not identified**.

The tempting next step: label individuals by $\hat{\tau}(x)$

$\hat{\tau}$ passed Part 3. Can we now say *who* is harmed, and *what* each patient gains? Three natural choices:

1. **Count the harmed:** the proportion of patients with $\hat{\tau}(x) < 0$.
2. **Pick whom to treat:** everyone with $\hat{\tau}(x) > c$.
3. **Tell a patient:** “your effect is $\hat{\tau}(x)$.”

People differ within x : $\Delta_i = Y_i(1) - Y_i(0) = \tau(X_i) + U_i$, $\mathbb{E}[U_i | X_i] = 0$, U_i never observed.

- Choice 2 survives: a rule is judged by its *mean* gain, $\mathbb{E}[\Delta_i \mathbf{1}\{\tau(X_i) > c\}] = \mathbb{E}[\tau(X) \mathbf{1}\{\tau(X) > c\}]$, and U_i averages out (Lecture 6).
- Choices 1 and 3 need the spread of U_i within x , e.g. $\mathbb{P}(\Delta_i < 0 | x) = \mathbb{P}\{U_i < -\tau(x) | x\}$.

We see one of $Y_i(1)$, $Y_i(0)$, never both: that spread is **not identified**.

$\tau(x) < 0$: the *average* person at x is harmed, not who, nor how many. **Q1** bounds how many; **Q2** gives each an interval.

The fraction harmed is not in the data

Q1 (partial identification): what proportion of patients does the treatment hurt?

Outcome Y : survive past 10 years (1) or not (0). Target $H = \mathbb{P}\{Y(0) = 1, Y(1) = 0\}$; the data give only the margins $p_a = \mathbb{P}\{Y(a) = 1\}$.

The fraction harmed is not in the data

Q1 (partial identification): what proportion of patients does the treatment hurt?

Outcome Y : survive past 10 years (1) or not (0). Target $H = \mathbb{P}\{Y(0) = 1, Y(1) = 0\}$; the data give only the margins $p_a = \mathbb{P}\{Y(a) = 1\}$. Two worlds, same data:

World A			
	$Y(1) = 1$	$Y(1) = 0$	total
$Y(0) = 1$	0.4	0.0	0.4
$Y(0) = 0$	0.2	0.4	0.6
total	0.6	0.4	1

harmed: 0%

World B			
	$Y(1) = 1$	$Y(1) = 0$	total
$Y(0) = 1$	0.2	0.2	0.4
$Y(0) = 0$	0.4	0.2	0.6
total	0.6	0.4	1

harmed: 20%

The fraction harmed is not in the data

Q1 (partial identification): what proportion of patients does the treatment hurt?

Outcome Y : survive past 10 years (1) or not (0). Target $H = \mathbb{P}\{Y(0) = 1, Y(1) = 0\}$; the data give only the margins $p_a = \mathbb{P}\{Y(a) = 1\}$. Two worlds, same data:

World A			
	$Y(1) = 1$	$Y(1) = 0$	total
$Y(0) = 1$	0.4	0.0	0.4
$Y(0) = 0$	0.2	0.4	0.6
total	0.6	0.4	1

harmed: 0%

World B			
	$Y(1) = 1$	$Y(1) = 0$	total
$Y(0) = 1$	0.2	0.2	0.4
$Y(0) = 0$	0.4	0.2	0.6
total	0.6	0.4	1

harmed: 20%

Both have margins $p_0 = 0.4$ (row total) and $p_1 = 0.6$ (column total).
Observation data only reveal marginal distribution.

What the margins allow: Fréchet–Hoeffding bounds

For each x the data give two numbers, $p_0(x) = \mathbb{P}\{Y(0) = 1 \mid x\}$ and $p_1(x) = \mathbb{P}\{Y(1) = 1 \mid x\}$. Harm is the cell $H(x) = \mathbb{P}\{Y(0) = 1, Y(1) = 0 \mid x\}$:

$$\max\{0, p_0(x) - p_1(x)\} \leq H(x) \leq \min\{p_0(x), 1 - p_1(x)\}.$$

What the margins allow: Fréchet–Hoeffding bounds

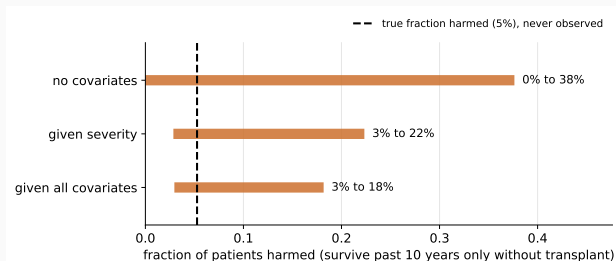
For each x the data give two numbers, $p_0(x) = \mathbb{P}\{Y(0) = 1 \mid x\}$ and $p_1(x) = \mathbb{P}\{Y(1) = 1 \mid x\}$. Harm is the cell $H(x) = \mathbb{P}\{Y(0) = 1, Y(1) = 0 \mid x\}$:

$$\max\{0, p_0(x) - p_1(x)\} \leq H(x) \leq \min\{p_0(x), 1 - p_1(x)\}.$$

- **Lower bound.** If fewer survive treated than untreated, someone was harmed: at least $p_0 - p_1$ of them.
- **Upper bound.** A harmed patient survives untreated (at most p_0) and dies treated (at most $1 - p_1$).
- **Then average over x .** Plug in $\hat{\mu}_0, \hat{\mu}_1$ for p_0, p_1 . More covariates, tighter bounds.

Sharp bounds on the fraction harmed, Kallus (2022).

The bounds, on the constructed registry



- No covariates: $H \in [0, 38\%]$.
Given severity: $[3\%, 22\%]$.
Given all covariates: $[3\%, 18\%]$.
- The truth, 5%, is never seen; every interval contains it.
- Here the bounds use the known margins. Estimating them well is the **open problem**.

Fréchet-Hoeffding bounds within X , Kallus (2022). Constructed registry with an unmeasured individual modifier.

An interval for one person's effect

Q2 (conformal inference): what will the treatment do for *this* patient?

Target $\Delta_i = Y_i(1) - Y_i(0)$. We want an interval $C(X_i)$ that covers it for 90% of patients:

$$\mathbb{P}\{\Delta_i \in C(X_i)\} \geq 0.90, \quad \text{probability over **patients**, not for a fixed patient.}$$

An interval for one person's effect

Q2 (conformal inference): what will the treatment do for *this* patient?

Target $\Delta_i = Y_i(1) - Y_i(0)$. We want an interval $C(X_i)$ that covers it for 90% of patients:

$$\mathbb{P}\{\Delta_i \in C(X_i)\} \geq 0.90, \quad \text{probability over \textbf{patients}, not for a fixed patient.}$$

1. **Interval for $Y_i(1)$:** $\hat{\mu}_1(X_i) \pm$ the 95% quantile of $|Y - \hat{\mu}_1(X)|$ on held-out **treated** rows, reweighted by $1/\hat{e}$ so they stand for everyone.
2. **Interval for $Y_i(0)$:** the same on **control** rows, weights $1/(1 - \hat{e})$.
3. **Subtract end to end:** $C = [L_1 - U_0, U_1 - L_0]$. Each interval misses at most 5% of patients, so C misses at most 10%.

An interval for one person's effect

Q2 (conformal inference): what will the treatment do for *this* patient?

Target $\Delta_i = Y_i(1) - Y_i(0)$. We want an interval $C(X_i)$ that covers it for 90% of patients:

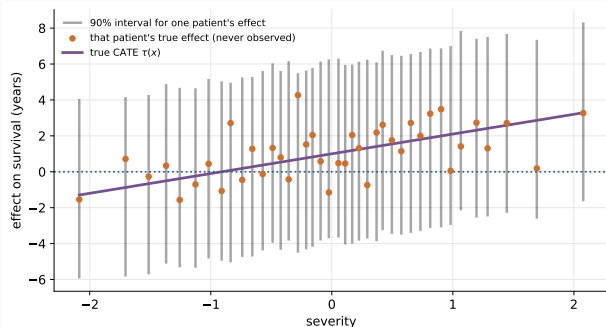
$$\mathbb{P}\{\Delta_i \in C(X_i)\} \geq 0.90, \quad \text{probability over \textbf{patients}, not for a fixed patient.}$$

1. **Interval for $Y_i(1)$:** $\hat{\mu}_1(X_i) \pm$ the 95% quantile of $|Y - \hat{\mu}_1(X)|$ on held-out **treated** rows, reweighted by $1/\hat{e}$ so they stand for everyone.
2. **Interval for $Y_i(0)$:** the same on **control** rows, weights $1/(1 - \hat{e})$.
3. **Subtract end to end:** $C = [L_1 - U_0, U_1 - L_0]$. Each interval misses at most 5% of patients, so C misses at most 10%.

Wide by construction: it carries both arms' noise plus the spread of U_i . The two arms share baseline noise, so it is also conservative.

Weighted split conformal for counterfactuals, Lei and Candès (2021); their nested method narrows the gap.

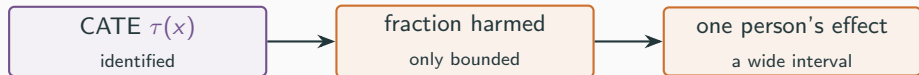
The intervals, on the constructed registry



- Every interval is about 10 years wide and every one contains 0.
- Coverage 100% against 90% nominal: valid but conservative.
- For scale: the true CATE has SD 1.1 years across patients.

Varying-effect registry with an unmeasured individual modifier $U_i \sim N(0, 1)$; $\alpha = 0.10$.

Three quantities, three kinds of answer

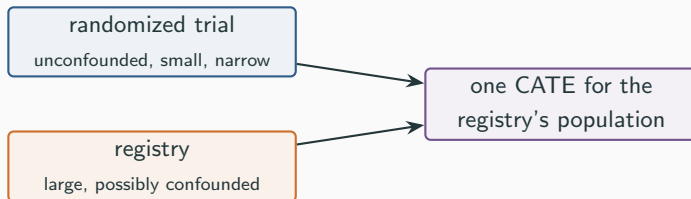


The closer the question gets to one person, the less the data can say.

Beyond one dataset

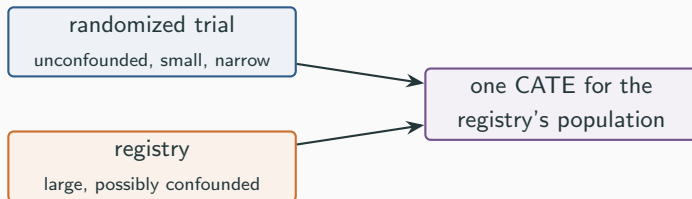
A small trial and a big registry

Each dataset has what the other lacks. Two questions, two directions of help:



A small trial and a big registry

Each dataset has what the other lacks. Two questions, two directions of help:



Q1 data fusion the trial removes the registry's hidden bias, through a **simple correction**.

Q2 transport the registry's covariates re-aim the trial's effect at the **right population**.

Kallus et al. (2018); review of the whole family: Colnet et al. (2024).

Why the registry's CATE can be biased

Q1 (data fusion): can a small trial fix a confounded registry?

The registry can only estimate the observable contrast

$$\begin{aligned}\omega(x) &= \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x], \\ \delta_a(x) &= \mathbb{E}[Y(a) \mid A = 1, x] - \mathbb{E}[Y(a) \mid A = 0, x], \quad a = 0, 1, \\ \omega(x) - \tau(x) &= e(x) \delta_0(x) + \{1 - e(x)\} \delta_1(x), \quad e(x) = \mathbb{P}(A = 1 \mid x)\end{aligned}$$

Why the registry's CATE can be biased

Q1 (data fusion): can a small trial fix a confounded registry?

The registry can only estimate the observable contrast

$$\begin{aligned}\omega(x) &= \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x], \\ \delta_a(x) &= \mathbb{E}[Y(a) \mid A = 1, x] - \mathbb{E}[Y(a) \mid A = 0, x], \quad a = 0, 1, \\ \omega(x) - \tau(x) &= e(x) \delta_0(x) + \{1 - e(x)\} \delta_1(x), \quad e(x) = \mathbb{P}(A = 1 \mid x)\end{aligned}$$

- $\delta_a(x)$: how the treated and controls at x differ in the *same* potential outcome $Y(a)$. Both are zero under no unmeasured confounding; the registry cannot check it.

Why the registry's CATE can be biased

Q1 (data fusion): can a small trial fix a confounded registry?

The registry can only estimate the observable contrast

$$\begin{aligned}\omega(x) &= \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x], \\ \delta_a(x) &= \mathbb{E}[Y(a) \mid A = 1, x] - \mathbb{E}[Y(a) \mid A = 0, x], \quad a = 0, 1, \\ \omega(x) - \tau(x) &= e(x) \delta_0(x) + \{1 - e(x)\} \delta_1(x), \quad e(x) = \mathbb{P}(A = 1 \mid x)\end{aligned}$$

- $\delta_a(x)$: how the treated and controls at x differ in the *same* potential outcome $Y(a)$. Both are zero under no unmeasured confounding; the registry cannot check it.
- A hidden *fitness* factor H makes a patient more likely to be transplanted *and* to live longer anyway.

Why the registry's CATE can be biased

Q1 (data fusion): can a small trial fix a confounded registry?

The registry can only estimate the observable contrast

$$\begin{aligned}\omega(x) &= \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x], \\ \delta_a(x) &= \mathbb{E}[Y(a) \mid A = 1, x] - \mathbb{E}[Y(a) \mid A = 0, x], \quad a = 0, 1, \\ \omega(x) - \tau(x) &= e(x) \delta_0(x) + \{1 - e(x)\} \delta_1(x), \quad e(x) = \mathbb{P}(A = 1 \mid x)\end{aligned}$$

- $\delta_a(x)$: how the treated and controls at x differ in the *same* potential outcome $Y(a)$. Both are zero under no unmeasured confounding; the registry cannot check it.
- A hidden *fitness* factor H makes a patient more likely to be transplanted *and* to live longer anyway. Then both $\delta_a > 0$, and $\omega(x) = \tau(x) + \text{bias}(x)$ with $\text{bias} > 0$.
- In the **trial**, randomization makes A independent of H : the gap is zero by design.

Use the trial to correct the registry

In the trial, assignment probability π is known, so a transformed outcome is unbiased for τ :

$$q_i = \{Y_i - g(X_i)\} \left\{ \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right\}, \quad \mathbb{E}[q_i \mid X_i = x] = \tau(x) \text{ for any fixed } g.$$

Use the trial to correct the registry

In the trial, assignment probability π is known, so a transformed outcome is unbiased for τ :

$$q_i = \{Y_i - g(X_i)\} \left\{ \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right\}, \quad \mathbb{E}[q_i \mid X_i = x] = \tau(x) \text{ for any fixed } g.$$

Using ω , we can estimate the bias: $\mathbb{E}[q - \omega(X) \mid x] = -\text{bias}(x)$.

Use the trial to correct the registry

In the trial, assignment probability π is known, so a transformed outcome is unbiased for τ :

$$q_i = \{Y_i - g(X_i)\} \left\{ \frac{A_i}{\pi} - \frac{1 - A_i}{1 - \pi} \right\}, \quad \mathbb{E}[q_i \mid X_i = x] = \tau(x) \text{ for any fixed } g.$$

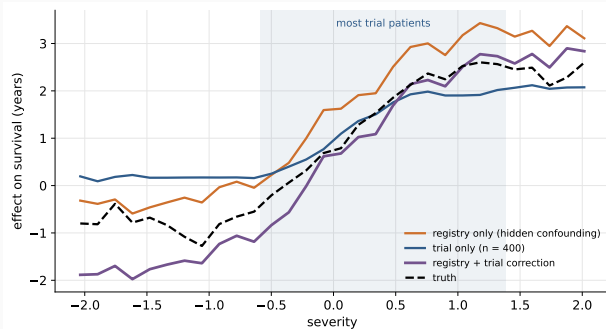
Using ω , we can estimate the bias: $\mathbb{E}[q - \omega(X) \mid x] = -\text{bias}(x)$.

1. **Registry:** fit $\hat{\omega}$ with Part 2's machinery.
2. **Trial:** regress $q_i - \hat{\omega}(X_i)$ on X_i with a **simple** model, here linear in severity and age.
3. **Combine:** $\hat{\tau}(x) = \hat{\omega}(x) + \widehat{\text{bias}}(x)$.

The bet: the bias is *simpler* than τ , so 400 trial patients can learn it although they could not learn τ . If there is no hidden confounding, $\widehat{\text{bias}} \approx 0$ and little is lost.

Formally: combining experimental and observational data, Kallus et al. (2018). Here $\pi = 0.5$ and g is the registry's fit of survival on X .

The correction, on the constructed registry



- RMSE over the registry: registry only 0.91, trial only 1.16, **combined 0.75** years.
- Registry alone: right shape, offset +0.64. Trial alone: unbiased but too noisy for a curve that bends.
- Outside the shaded range every trial-based curve is **extrapolating**.

Constructed registry with a hidden fitness factor that raises both transplant and survival; trial $n = 400$.

Moving an effect to a new population

Q2 (transport): the trial enrolled sicker patients; what is the effect in the registry?

The trial's ATE averages $\tau(x)$ over *its* patients. The registry wants the same $\tau(x)$ averaged over *its* own:

$$\tau_{\text{trial}} = \mathbb{E}_{\text{trial}}\{\tau(X)\}, \quad \tau_{\text{registry}} = \mathbb{E}_{\text{registry}}\{\tau(X)\}, \quad \text{same } \tau(x), \text{ different } X.$$

Moving an effect to a new population

Q2 (transport): the trial enrolled sicker patients; what is the effect in the registry?

The trial's ATE averages $\tau(x)$ over *its* patients. The registry wants the same $\tau(x)$ averaged over *its* own:

$$\tau_{\text{trial}} = \mathbb{E}_{\text{trial}}\{\tau(X)\}, \quad \tau_{\text{registry}} = \mathbb{E}_{\text{registry}}\{\tau(X)\}, \quad \text{same } \tau(x), \text{ different } X.$$

1. **Fit** $\hat{\tau}(x)$ in the trial, where it is unconfounded.
2. **Average** it over the registry's covariates, $\frac{1}{n} \sum_{i \in \text{registry}} \hat{\tau}(X_i)$; or, equivalently, reweight trial patients to look like the registry.
3. **Check support:** registry patients outside the trial's range of X are extrapolation again.

Needs: $\tau(x)$ is the same function in both populations, so every effect modifier is in X .

Formally: transportability and generalizability, Colnet et al. (2024).

The shift, on the constructed registry

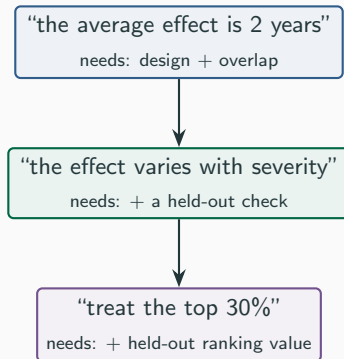


- Trial patients are sicker. Reweighting them to the registry's covariates (black outline) moves toward the target, but cannot reach severities the trial never enrolled.
- Trial ATE as run 1.46; reweighted 1.14; averaging the trial CATE over the registry 1.03. Registry truth 0.80.

Same constructed trial and registry as the previous slide.

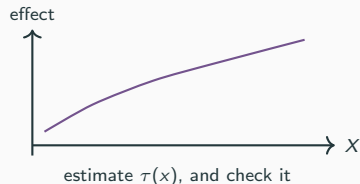
Wrap-up

Three claims, three burdens of proof

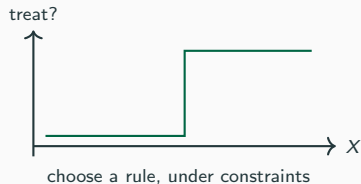


Next: from an estimated effect to a decision

Today



Lecture 6



Sequential decisions, offline policy learning, and time-varying treatments (Athey and Wager, 2021).

Project proposals are due today, October 2.

References

- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.
- Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.
- Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2025). Fisher–Schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. *Econometrica*, 93(4):1121–1164.
- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. Online book.
- Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191.
- Curth, A. and van der Schaar, M. (2023). In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6623–6642.

References and further reading ii

- Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.
- Kallus, N. (2022). What's the harm? Sharp bounds on the fraction negatively affected by treatment. In *Advances in Neural Information Processing Systems 35*, pages 15996–16009.
- Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems 31*.
- Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.
- Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):911–938.
- Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.
- Radcliffe, N. J. and Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. *Stochastic Solutions White Paper*.

- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954.
- Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.
- van der Laan, M. J. (2006). Statistical inference for variable importance. Working Paper 188, U.C. Berkeley Division of Biostatistics.
- van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1). Article 11.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2025). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51.

Questions?