Combining Observational and Experimental Data

Robin Evans
University of Oxford

JICI Workshop, UC Berkeley
7th September 2022

DEPARTMENT OF
STATISTICS

Collaborators

Xi Lin, University of Oxford

Outline

Randomized Trials vs Observational Studies

Randomized trials

- give unbiased estimates for causal effects;

but

- they are expensive;
- sample sizes may be small;
- typically have exclusion criteria.

Observational studies

- may better represent the target population;
- often have much larger sample sizes;
- important subgroups may be much better represented;

but

- they are not randomized!

Set Up

Suppose we have two datasets:

- $\mathcal{D}_e$, with sample size n_e, from a **randomized controlled trial**; and
- $\mathcal{D}_o$, with sample size $n_o \gg n_e$, some **observational database**.

For simplicity, assume they both have i.i.d. observations (from P_e, P_o) of $X = (\boldsymbol{Z}, T, Y)$ where:

- T is a treatment;
- Y is an outcome;
- $\boldsymbol{Z}$ is a collection of confounders/effect modifiers.

Target Parameter and Loss Function

The RCT $\mathcal{D}_e$ is a 'gold standard' dataset, but we would like to be able to use $\mathcal{D}_o$ to improve our inference about certain subgroups.

That is, we want to optimally estimate the **conditional average treatment effect** (CATE) using these two datasets:

$$\text{CATE}(\boldsymbol{z}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{Z} = \boldsymbol{z}].$$

We assume that $Y(t) \mid \boldsymbol{Z}$ has the **same distribution** under P_e and P_o.

Given a target parameter θ, our loss function is the **mean squared error**:

$$\text{MSE} = \mathbb{E}\|\hat{\theta} - \theta\|^2 = \text{Var}\,\hat{\theta} + (\text{Bias}\,\hat{\theta})^2.$$

The randomized trial data has zero bias, but high variance;
the observational data has lower variance, but may be confounded.

So there is naturally a **bias-variance tradeoff**.

Some Solutions

1. **Minimize Mean Squared Error**. Find the convex combination that minimizes the mean squared error of the pooled estimate (Oberst et al., 2022).

2. **Shrinkage.** Use James-Stein approach to shrink RCT estimate towards observational. Guaranteed to reduce the overall MSE (Green and Strawderman, 1991; Rosenman et al., 2020).

3. **Experimental Grounding**. Assume a parametric model can be used to correct for confounding in the observational dataset (Kallus et al., 2018).

4. **Power Likelihood.** Take the joint likelihood but raise it to a power $\eta \leq 1$ for the observational data; if chosen correctly this will give good inference, even if there is unobserved confounding (e.g. Holmes and Walker, 2017).

Solution 1: Minimize MSE

Assume that $\theta_o = \theta_e + \delta$ for some unknown bias $\delta \in \mathbb{R}^d$.

Then we want to pick $\lambda \in [0, 1]$ such that

$$\hat{\theta}_\lambda = \lambda \hat{\theta}_o + (1 - \lambda)\hat{\theta}_e$$

has the smallest possible MSE.

As $\lambda \to 1$ the bias increases, as $\lambda \to 0$ the variance increases.

The optimal λ is given by

$$\lambda = \frac{\sigma_e^2}{\|\delta\|^2 + \sigma_o^2 + \sigma_e^2}.$$

Different Solutions

Oberst et al. (2022) suggest the following (plug-in) estimator:

$$\hat{\lambda}_{\text{ober}} = \frac{\hat{\sigma}_e^2}{\|\hat{\theta}_e - \hat{\theta}_o\|^2 + \hat{\sigma}_o^2 + \hat{\sigma}_e^2};$$

Rosenman et al. (2020) note that $\mathbb{E}\|\hat{\theta}_e - \hat{\theta}_o\|^2$ is equal to the denominator, and use

$$\hat{\lambda}_{\text{JS}} = \frac{\hat{\sigma}_e^2}{\|\hat{\theta}_e - \hat{\theta}_o\|^2}.$$

Oberst et al. conclude that neither method is always better than the other.

Solution 2: Shrinkage

Stein's Paradox

Let $W_i \sim N(\mu_i, 1)$ independently for $i = 1, \ldots, d$, with $d \geq 3$.
Then $(W_1, \ldots, W_d)$ is inadmissible for $(\mu_1, \ldots, \mu_d)$.

The proof is to show that

$$\tilde{\mu}_i = \left(1 - \frac{d-2}{\|\boldsymbol{W}\|}\right) W_i$$

has a smaller MSE than $\hat{\mu}_i = W_i$ does!

This is a remarkable seeming result, but is really just due to the stabilization given by **shrinking** the observation towards zero.

Shrinkage Method

Stratify experimental data, so that parameter estimates between the groups are independent.

Then estimators of CATE parameters are a (scaled) standard normal vector.

This means we have:

$$\begin{aligned} \hat{\theta}_e, &\quad \text{unbiased estimate of } \theta \\ \hat{\theta}_o, &\quad \text{(possibly) biased estimate of } \theta, \end{aligned}$$

each with diagonal covariance matrix.

Then use James-Stein type approach to shrink the experimental estimate towards the observational one.

This is guaranteed to give a smaller overall **mean squared error**. (Green and Strawderman, 1991; Rosenman et al. 2020).

Obvious disadvantage is that result is **not true** if components of $\hat{\theta}_e$ are dependent.

Shrinkage Method

The shrinkage estimator is

$$\hat{\theta}_{\text{shr}} = \hat{\theta}_o + \left(1 - \sigma_e^2 \frac{d-2}{\|\hat{\theta}_e - \hat{\theta}_o\|^2}\right)_+ (\hat{\theta}_e - \hat{\theta}_o),$$

which is essentially shrinking $\hat{\theta}_e$ towards $\hat{\theta}_o$.
(Here σ_e^2 is the variance of each component of $\hat{\theta}_e$.)

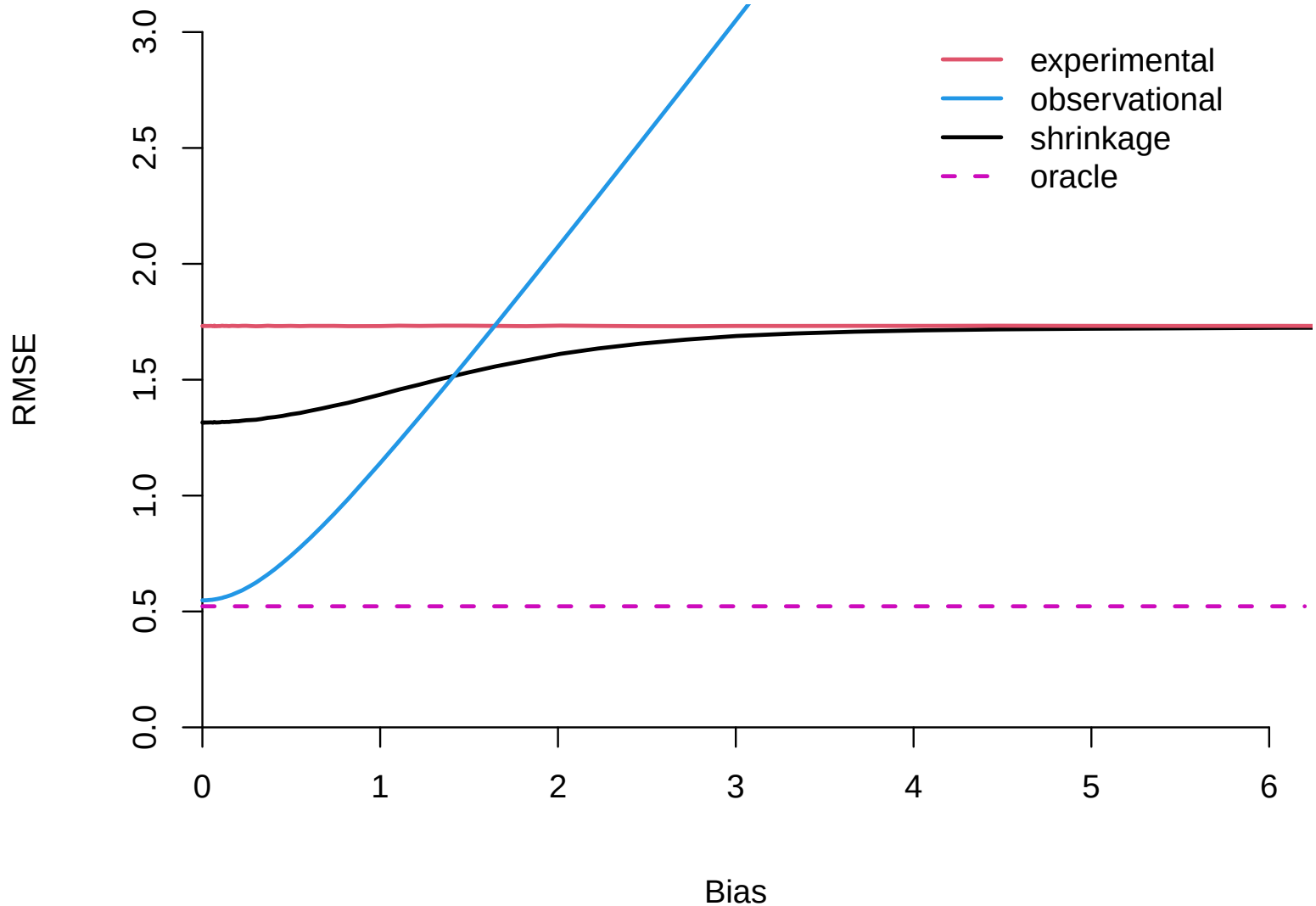
We perform a simulation for $d = 3$ where we assume

$$10 \operatorname{Var} \hat{\theta}_o = \operatorname{Var} \hat{\theta}_e,$$

and that the bias is of the form $(\delta_1, 0, 0)^T$.

We take $\sigma_e^2 = 1$ as known (but this is easy to estimate).

James-Stein Simulations



Solution 3: Experimental Grounding

Kallus et al. (2018) use **experimental grounding**, which assumes that there is a parametric function φ that explains the bias due to unobserved confounders:

$$\varphi(\boldsymbol{z}) = \{E[Y(1) \mid \boldsymbol{Z} = \boldsymbol{z}] - E[Y(1) \mid \boldsymbol{Z} = \boldsymbol{z}, T = 1]\} + \\ - \{E[Y(0) \mid \boldsymbol{Z} = \boldsymbol{z}] - E[Y(0) \mid \boldsymbol{Z} = \boldsymbol{z}, T = 0]\}.$$

They then attempt to learn φ by comparing predictions using experimental and observational datasets.

Estimation of the parameters is by least squares.

They demonstrate the methodology on the STAR Dataset (an RCT) by artificially inducing confounding (see later on).

Solution 4: Power Likelihood

Weight the observational data at a lower level than the experimental (say $\eta < 1$), and then perform likelihood-based inference.

How do we choose η?

One approach is to maximize the **expected log pointwise predictive density** (ELPD):

$$\text{ELPD}(\eta) = \mathbb{E}_X \log p_\eta(X \mid \boldsymbol{x}),$$

where $p_\eta(X \mid \boldsymbol{x})$ is the posterior predictive when using the power η, and the expectation is with respect to the 'true' distribution.

Estimating η

We can approximate the ELPD using the **widely applicable information criterion** (WAIC) of Watanbe (2010).

Simply use the ordinary posterior to estimate the density of each observation, and subtract the WAIC.

$$\widehat{\mathrm{ELPD}}(\eta) = \frac{1}{n_e} \sum_{i=1}^{n_e} \log \hat{p}_\eta(x_i \mid \boldsymbol{x}) - \hat{d}_{\mathrm{WAIC}},$$

where $\hat{p}_\eta(x_i \mid \boldsymbol{x})$ is estimated using an MC sample of the parameters.

Computationally intensive to evaluate, as we need to estimate function above over a **grid** of values for η; may be OK if we only do it once.

If η is more complicated (e.g. a two-dimensional parameter) then this quickly becomes a problem (though maybe VI can help).

Likelihood

A more basic question: how do we even get a likelihood for the observational data?

Answer

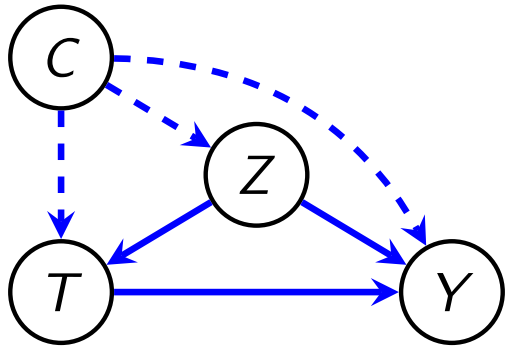
We can use the **frugal parameterization** (E. and Didelez, 2021).

Allows us to have a **parametric** representation of various causal quantites.

Frugal Parameterization Summary

- Describe the parametric distribution **after** the relevant intervention, using (e.g.) a copula to join outcome and anything in $\boldsymbol{Z}$;
- then reweight to obtain the 'observational' distribution;
- can simulate using rejection sampling.

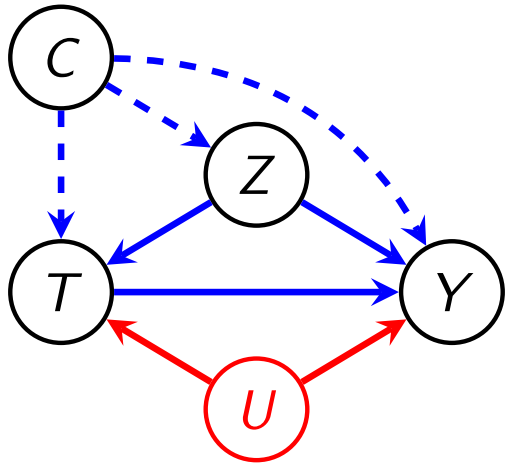
Frugal Parameterization



For our problem, separately specify (nice, parametric) models for:

- $p(c, z, t)$; ('the past')
- $p(y(t) \mid c)$; (quantity of interest)
- $\phi^*_{ZY|CT}$. (some dependence measure)

Frugal Parameterization (with misspecification)



Separately specify (nice, parametric) models for:

- $p(u, c, z, t)$; ('the past')
- $p(y(t) \mid u, c)$; (quantity of interest)
- $\phi^*_{ZY|CTU}$. (some dependence measure)

Simulation

Consider the following observational setup:

$$U, C \sim \text{Bernoulli}(0.5) \qquad Z \mid C \sim N(\mu_z, 1)$$
$$T \mid C, Z, U \sim \text{Bernoulli}(\mu_t) \qquad Y(t) \mid C \sim N(\mu_y, 1).$$

where

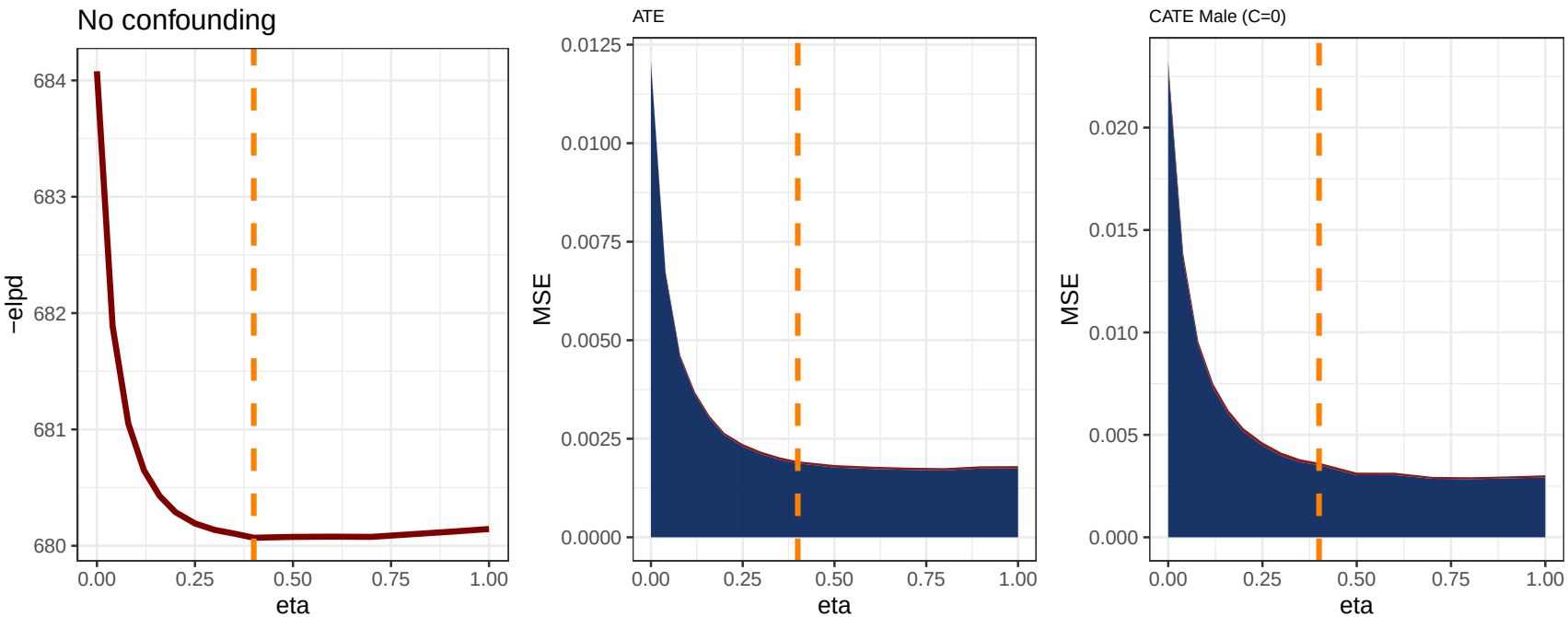
$$\begin{aligned} \mu_z &= 0.2 + 0.6\, C \\ \text{logit}\, \mu_t &= 0.5 + 0.1\, C + 0.6\, Z + 0.4\, C\, Z + \gamma U \\ \mu_y &= 0.6 + 0.2\, C \qquad\qquad + 1.1\, C\, T + \gamma U \end{aligned}$$

Gaussian copula w. correlation $2\,\text{expit}(1) - 1 \approx 0.462$ between Y and Z.

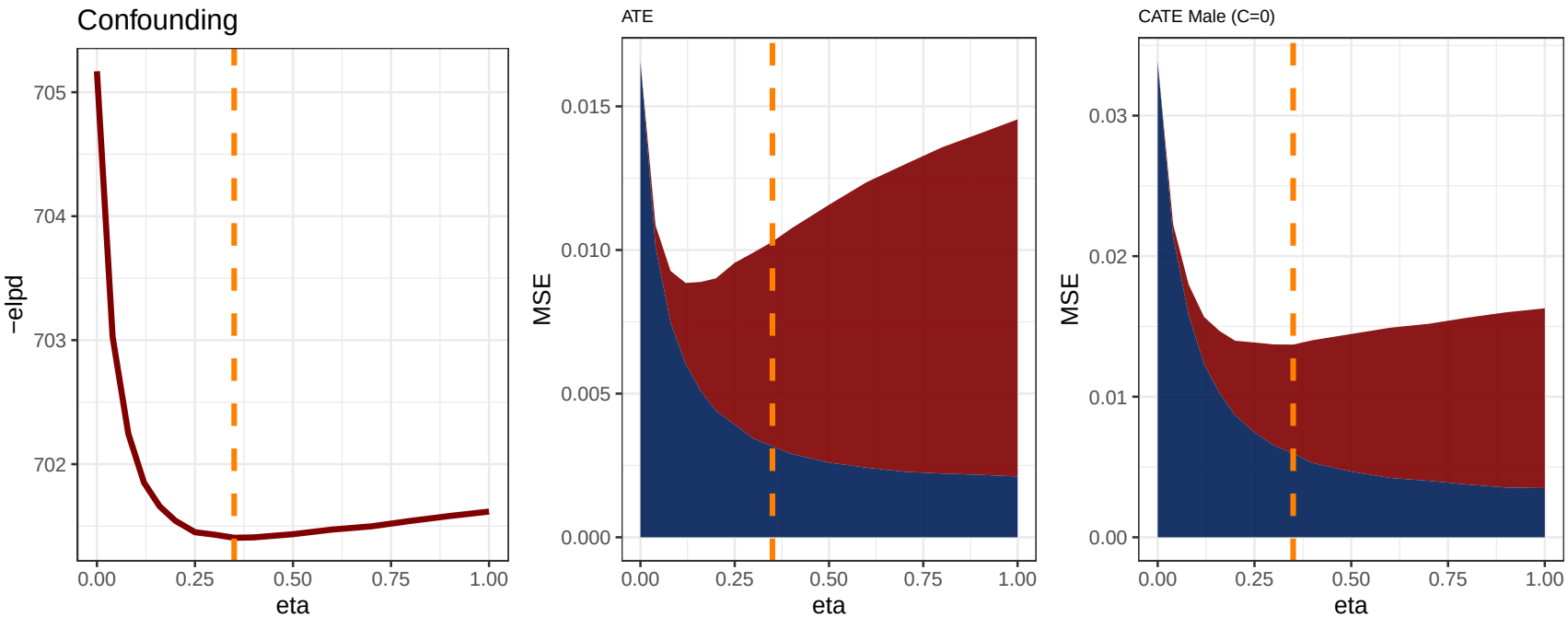
For the experimental data, we just take $\mu_t = 0.5$.

Suppose we have sample sizes of $n_e = 250$ randomized individuals and $n_o = 2{,}500$ units in the observational study.

Results (no confounding: $\gamma = 0$)



Results ($\gamma = 0.75$)



STAR Data

Inspired by Kallus et al. (2018), we construct our datasets from Tennessee's Student Teacher Achievement Ratio Study, an RCT.

Over 7,000 students in 79 schools were randomly assigned into one of three interventions:

- small class (13 to 17 students per teacher);
- regular class (22 to 25 students per teacher); and
- regular-with-aide class (22 to 25 students with a full-time aide)

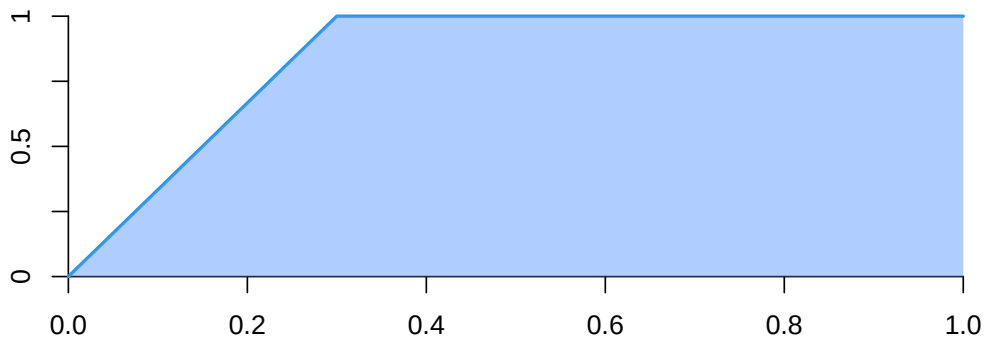
To get a confounded observational dataset:

- take a variable U (school type) that predicts the outcome strongly;
- pick a subset of values for U to obtain an unconfounded dataset;
- condition upon T, U and Y to to make remaining data confounded;
- marginalize U.

Now there is **unobserved confounding** in the selected subset.

Details

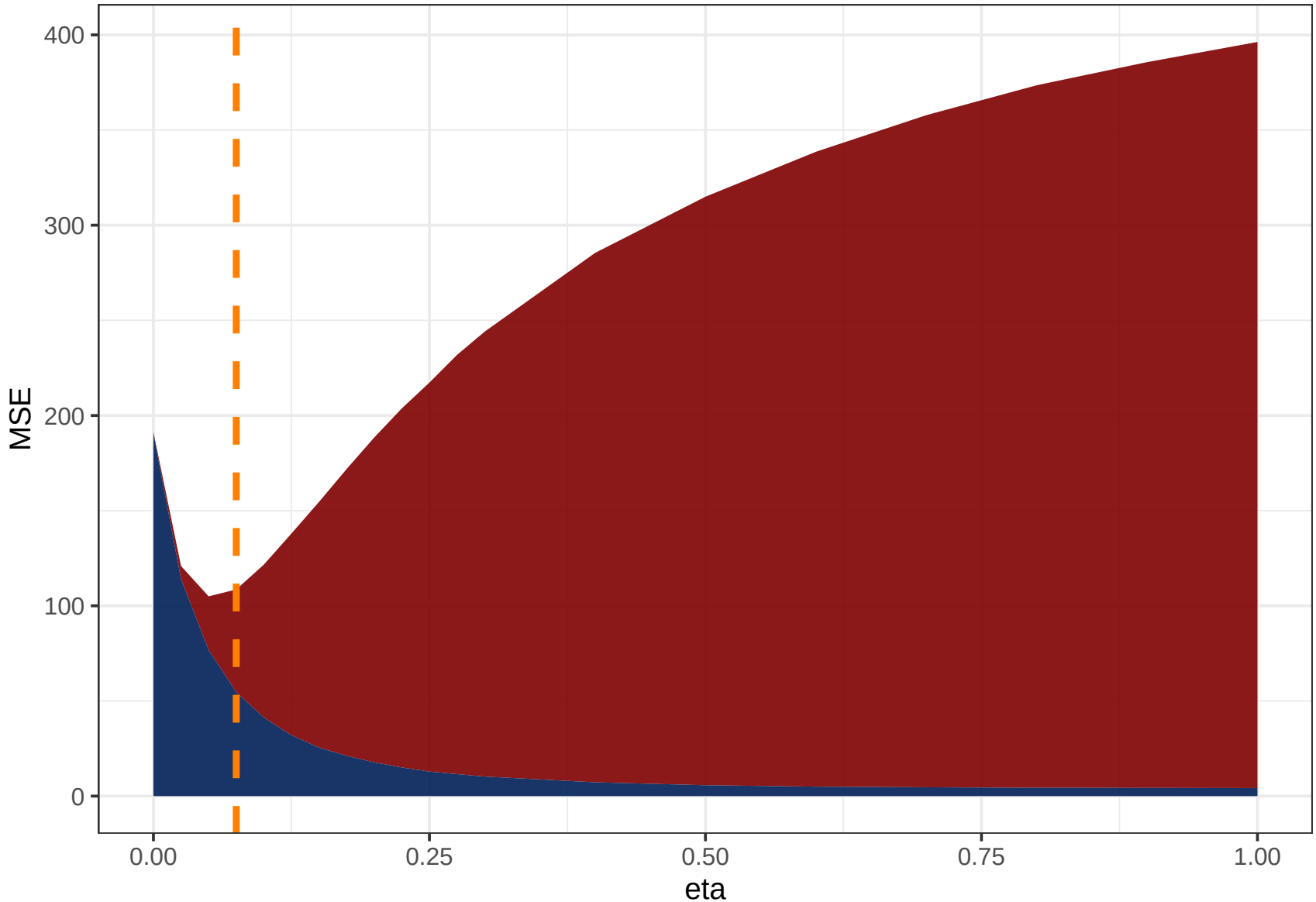
1. Take all treated individuals ($T = 1$).
2. Those with an outcome below the 30th percentile are **down-weighted** in proportion to their quantile. (e.g. 10th percentile score has 1/3 chance of score at 30th percentile or above.)



3. Then select 1,000 observations using this weighting for the confounded data.

The confounding means we obtain a naïve estimate of 57.0 rather than the original dataset's 38.4.

Results



Summary

- Experimental data has better **internal validity**, observational data has better **external validity**.
- Combining RCTs with observational databases will 'clearly' lead to improved causal inference, especially for **small groups**.

- There are various approaches to doing this:
  - using shrinkage;
  - by experimental grounding;
  - minimizing the MSE directly;
  - using a power likelihood.

- Estimating the right parameter(s) for combining these datasets is still a **big challenge**!

Thank you!

References

Evans and Didelez. Parameterizing and simulating from causal models. *arXiv preprint* 2109.03694, 2021.

Green and Strawderman. A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators, *JASA*, 1991.

Holmes and Walker. Assigning a value to a power likelihood in a general Bayesian model, *Biometrika*, 2017.

Kallus et al. Removing Hidden Confounding by Experimental Grounding, *NeurIPS*, 2018.

Oberst et al. Bias-robust Integration of Observational and Experimental Estimators, *arXiv preprint* 2205.10467, 2022.

Rosenman et al. Combining Observational and Experimental Datasets Using Shrinkage Estimators, *arXiv preprint* 2002.06708, 2020.