# Shrinkage Estimation for Causal Inference and Experimental Design

**Evan T. R. Rosenman**†, Guillaume Basse, Mike Baiocchi, Art B. Owen, Francesca Dominici, and Luke Miratrix

† Assistant Professor, Claremont McKenna College

September 16, 2023

# Motivating Setting

**Randomized Controlled Trials (RCT)**

- Researcher controls assignment to treatment
  - Relatively few assumptions for unbiasedness
  - Often costly, small
- "Unbiased but imprecise"

**Observational Databases**

- Treatment assignments observed, but not controlled
  - Confounding $\implies$ unverifiable assumptions for unbiasedness
  - Large, often inexpensive.
- "Precise, but biased"

## Our Approach

We consider how to...

- **design shrinkage estimators to merge observational and RCT data** $\rightarrow$ two paradigms!
- **improve experimental design using shrinkers?**

# Outline

# Central Role of Stratification

- Work in a stratified setting, with $K$ strata.
  - Characterize heterogeneity in treatment effect
  - Arise from subject matter expertise, modern ML method, etc.
- Each unit $i$ in RCT + ODB has associated stratum indicator $S_i \in \{1, \ldots, K\}$
- (Unobserved) Conditional avg. stratum treatment effects:

$$\tau_{rk} = \mathbb{E}_R \left( Y_i(1) - Y_i(0) \mid S_i = k \right)$$
$$\tau_{ok} = \mathbb{E}_O \left( Y_i(1) - Y_i(0) \mid S_i = k \right)$$

# Central Role of Stratification

- Work in a stratified setting, with $K$ strata.
    - Characterize heterogeneity in treatment effect
    - Arise from subject matter expertise, modern ML method, etc.
- Each unit $i$ in RCT + ODB has associated stratum indicator $S_i \in \{1, \ldots, K\}$
- (Unobserved) Conditional avg. stratum treatment effects:

$$\tau_{rk} = \mathbb{E}_R \left( Y_i(1) - Y_i(0) \mid S_i = k \right)$$
$$\tau_{ok} = \mathbb{E}_O \left( Y_i(1) - Y_i(0) \mid S_i = k \right)$$

**Transportability of CATEs**: For $k = 1, \ldots, K$, treatment effects $\tau_{ok} = \tau_{rk}$, and we call their common value $\tau_k$.
Define $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)^{\mathsf{T}}$.

# Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = \left(\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}\right), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right).$$

# Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = (\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = (\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}).$$

- Under mild conditions, we have

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \sim N(\boldsymbol{\tau}, \Sigma_r), \quad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \sim (\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o)$$

  for bias $\boldsymbol{\xi}$ and covariance matrices $\Sigma_r$ and $\Sigma_o$

  - $\boldsymbol{\xi}$ cannot be estimated from obs data alone

# Setup

- Collect our estimators into vectors:

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} = \left(\hat{\tau}_{r1}, \ldots, \hat{\tau}_{rK}\right), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right).$$

- Under mild conditions, we have

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \sim N\left(\boldsymbol{\tau}, \Sigma_r\right), \qquad \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \sim \left(\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o\right)$$

  for bias $\boldsymbol{\xi}$ and covariance matrices $\Sigma_r$ and $\Sigma_o$
    - $\boldsymbol{\xi}$ cannot be estimated from obs data alone
- Seek to design shrinkage estimator $\hat{\boldsymbol{\tau}} = f(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$ to minimize expected squared error loss,

$$\mathcal{L}(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \sum_k \left(\hat{\tau}_k - \tau_k\right)^2.$$

# Useful Prior Work

- **Shrinkage estimation**: a rich literature stretching back to multivariate normal mean estimation work of Stein (1956)
- Green and Strawderman (1991) and Green et al. (2005) propose estimators for shrinkage between ...
  - a normal, unbiased estimator (like $\hat{\tau}_r$), and
  - a biased estimator (like $\hat{\tau}_o$)

# Outline

1. Assumptions and Loss Function
2. Inference
   - Positing Shrinkage Structure
   - Using a Hierarchical Model
3. Application to the WHI
4. Experimental Design

# Outline

# A Recipe for Estimators

1. Posit a structure for the shrinkage estimator

$$f(\hat{\tau}_r, \hat{\tau}_o) = \hat{\tau}_r - \boldsymbol{g}(\hat{\tau}_r, \hat{\tau}_o)$$

for any differentiable $g$ satisfying $E(||\boldsymbol{g}||^2) < \infty$.

# A Recipe for Estimators

1. Posit a structure for the shrinkage estimator

$$f(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \boldsymbol{g}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

   for any differentiable $g$ satisfying $E(||\boldsymbol{g}||^2) < \infty$.

2. Following common precedent (Li et al., 1985; Xie et al., 2012), minimize unbiased risk estimate,

$$\text{URE} = \frac{1}{K}\left(\text{Tr}\left(\Sigma_r\right) + \sum_{k=1}^{K} g_k^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) - 2\sigma_{rk}^2 \frac{\partial g_k(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}{\hat{\tau}_{rk}}\right)$$

   over hyperparameters to obtain the estimator.

# Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage factor $\lambda$ across components. Denote generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \,.$$

# Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage factor $\lambda$ across components. Denote generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) .$$

Then, URE evaluates to

$$\mathrm{URE}(\lambda) = \mathrm{Tr}\left(\Sigma_r\right) + \lambda^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right) - 2\lambda \mathrm{Tr}(\Sigma_r)$$

## Case 1: Common Shrinkage Factor

We consider shrinkage estimators which share a common shrinkage factor $\lambda$ across components. Denote generic estimator as

$$\boldsymbol{\kappa}(\lambda, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \,.$$

Then, URE evaluates to

$$\mathrm{URE}(\lambda) = \mathrm{Tr}\left(\Sigma_r\right) + \lambda^2 \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right) - 2\lambda \mathrm{Tr}(\Sigma_r)$$

which has minimizer in $\lambda$,

$$\lambda_1^{\mathrm{URE}} = \frac{\mathrm{Tr}(\Sigma_r)}{\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)^{\mathsf{T}} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\right)} \,.$$

# Useful Properties of $\lambda_1^{\text{URE}}$ (I)

1. Define

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \lambda_1^{\text{URE}} \left( \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \right)$$

**Lemma ($\boldsymbol{\kappa}_1$ Risk Guarantee)**

*Suppose $4 \max_k \sigma_{rk}^2 < \sum_k \sigma_{rk}^2$. Then $\boldsymbol{\kappa}_1$ has risk strictly less than that of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$.*

- Requires a dimension of at least $K = 5$.
- May require substantially larger $K$ if high heteroscedasticity

# Useful Properties of $\lambda_1^{\mathrm{URE}}$ (II)

2. Its positive part analogue,

$$\boldsymbol{\kappa}_{1+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{\lambda_1^{\mathrm{URE}}\right\}_{[0,1]} \left(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}\right) ,$$

where

$$\{u\}_{[0,1]} = \min(\max(u, 0), 1) ,$$

satisfies the following notion of optimality:

# Useful Properties of $\lambda_1^{\mathrm{URE}}$ (III)

**Theorem ($\boldsymbol{\kappa}_{1+}$ Asymptotic Risk)**

*Suppose*

$$\limsup_{K\to\infty} \frac{1}{K}\sum_k \sigma_{rk}^2 \xi_k^2 < \infty\,, \quad \limsup_{K\to\infty} \frac{1}{K}\sum_k \sigma_{rk}^2 \sigma_{ok}^2 < \infty\,,$$

$$\text{and} \quad \limsup_{K\to\infty} \frac{1}{K}\sum_k \sigma_{rk}^4 < \infty\,.$$

*Then, in the limit $K \to \infty$, $\boldsymbol{\kappa}_{1+}$ has the lowest risk among all estimators with a shared shrinkage factor across components.*

# Case 2: Variance-Weighted Shrinkage Factor

This procedure is general purpose. For example, may instead want an estimator that shrinks each component proportionally to $\sigma^2_{rk}$.

Easy to solve for

$$\boldsymbol{\kappa}_2 = \boldsymbol{\kappa}(\lambda_2^{\text{URE}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \frac{\text{Tr}(\Sigma_r^2)\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}\Sigma_r^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

and its positive-part improvement,

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{\frac{\text{Tr}(\Sigma_r^2)\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}\Sigma_r^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})}\right\}_{[0,1]}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) \ .$$

# Outline

# Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

# Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

Simple model generalizing one introduced in Green and Strawderman (1991):

$$\boldsymbol{\tau} \sim \mathcal{N}\left(0, \eta^2 \boldsymbol{I}_K\right),$$
$$\boldsymbol{\xi} \sim \mathcal{N}\left(0, \gamma^2 \boldsymbol{I}_K\right),$$

# Alternative Approach: Hierarchical Model

- In prior section, functional form was **imposed** by the researcher based on problem parameters
- An alternative approach is to derive the functional form from a **hierarchical model**

Simple model generalizing one introduced in Green and Strawderman (1991):

$$\begin{aligned} \boldsymbol{\tau} &\sim \mathcal{N}\left(0, \eta^2 \boldsymbol{I}_K\right), \\ \boldsymbol{\xi} &\sim \mathcal{N}\left(0, \gamma^2 \boldsymbol{I}_K\right), \\ \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \mid \boldsymbol{\tau} &\sim \mathcal{N}\left(\boldsymbol{\tau}, \Sigma_r\right), \text{ and} \\ \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \mid \boldsymbol{\tau}, \boldsymbol{\xi} &\sim \mathcal{N}\left(\boldsymbol{\tau} + \boldsymbol{\xi}, \Sigma_o\right). \end{aligned} \tag{1}$$

for **unknown** hyperparameters $\eta^2$ and $\gamma^2$, but **known** covariance matrices $\Sigma_r, \Sigma_o$.

# Estimator Form

Estimator can be constructed as the **posterior mean** of $\boldsymbol{\tau}$ under this model, which evaluates to

$$\psi_k(\eta^2,\gamma^2) = \underbrace{\left(\frac{\eta^2\left(\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2\right)}{\sigma_{rk}^2\left(\gamma^2+\sigma_{ok}^2\right)+\eta^2\left(\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2\right)}\right)}_{\boldsymbol{a_k}(\boldsymbol{\eta}^2,\gamma^2)\text{: aggregate shrinkage toward zero}} \times \left(\underbrace{\frac{\left(\gamma^2+\sigma_{ok}^2\right)}{\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2}}_{\substack{\boldsymbol{\lambda_k}(\boldsymbol{\eta}^2,\gamma^2)\text{:}\\ \text{data-driven weight}}}\hat{\tau}_{rk} + \underbrace{\frac{\sigma_{rk}^2}{\gamma^2+\sigma_{ok}^2+\sigma_{rk}^2}}_{1-\boldsymbol{\lambda_k}(\boldsymbol{\eta}^2,\gamma^2)}\hat{\tau}_{ok}\right). \qquad (2)$$

# Estimator Form

Estimator can be constructed as the **posterior mean** of $\boldsymbol{\tau}$ under this model, which evaluates to

$$\psi_k(\eta^2, \gamma^2) = \underbrace{\left( \frac{\eta^2 \left(\gamma^2 + \sigma_{ok}^2 + \sigma_{rk}^2\right)}{\sigma_{rk}^2 \left(\gamma^2 + \sigma_{ok}^2\right) + \eta^2 \left(\gamma^2 + \sigma_{ok}^2 + \sigma_{rk}^2\right)} \right)}_{\boldsymbol{a_k}(\boldsymbol{\eta}^2, \boldsymbol{\gamma}^2)\text{: aggregate shrinkage toward zero}} \times \left( \underbrace{\frac{\left(\gamma^2 + \sigma_{ok}^2\right)}{\gamma^2 + \sigma_{ok}^2 + \sigma_{rk}^2}}_{\substack{\boldsymbol{\lambda_k}(\boldsymbol{\eta}^2, \boldsymbol{\gamma}^2)\text{:} \\ \text{data-driven weight}}} \hat{\tau}_{rk} + \underbrace{\frac{\sigma_{rk}^2}{\gamma^2 + \sigma_{ok}^2 + \sigma_{rk}^2}}_{1 - \boldsymbol{\lambda_k}(\boldsymbol{\eta}^2, \boldsymbol{\gamma}^2)} \hat{\tau}_{ok} \right). \tag{2}$$

This is the **double-shrinkage** property: take a data-driven convex combo of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ and then a Stein-like shrinakge toward zero.

# Versions of the Estimator (I)

To construct a usable estimator, need estimates of $\eta^2, \gamma^2$.
Use three approaches from Xie et al. (2012)

**Moment-Matching**: Observing that

$$\mathbb{E}\left(||\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}||_2^2\right) = \mathrm{Tr}(\Sigma_r) + K\eta^2, \qquad \text{and}$$
$$\mathbb{E}\left(||\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}||_2^2\right) = \mathrm{Tr}(\Sigma_o) + \mathrm{Tr}(\Sigma_r) + K\gamma^2,$$

use the estimates:

$$\hat{\eta}^2_{\mathrm{mm}} = \frac{1}{K}\left(||\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}||_2^2 - \mathrm{Tr}(\Sigma_r)\right)_+$$
$$\hat{\gamma}^2_{\mathrm{mm}} = \frac{1}{K}\left(||\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}||_2^2 - \mathrm{Tr}(\Sigma_r) - \mathrm{Tr}(\Sigma_o)\right)_+ .$$

# Versions of the Estimator (II)

**Maximum Likelihood**: Observing that

$$\mathcal{L}(\eta^2, \gamma^2) \propto \prod_k \left(\eta^2 + \sigma_{rk}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{rk}^2}{2\left(\eta^2+\sigma_{rk}^2\right)}} \times$$

$$\prod_k \left(\eta^2 + \gamma^2 + \sigma_{ok}^2\right)^{-1/2} e^{-\frac{\hat{\tau}_{ok}^2}{2\left(\eta^2+\gamma^2+\sigma_{ok}^2\right)}} .$$

We can numerically optimize to obtain the estimates

$$(\hat{\eta}^2_{\text{mle}}, \hat{\gamma}^2_{\text{mle}}) = \max_{\eta^2, \gamma^2 \geq 0} \log \left( \mathcal{L}(\eta^2, \gamma^2) \right) .$$

## Versions of the Estimator (III)

**URE Minimization**: We can use the same URE-minimization approach as in the prior section! Here,

$$\begin{aligned}\mathrm{URE}(\eta^2,\gamma^2) =& \mathrm{Tr}(\Sigma_r) + \sum_k \left(\psi_k(\eta^2,\gamma^2) - \hat{\tau}_{rk}\right)^2 - \\ & 2\sum_k \sigma_{rk}^2 \cdot \left(1 - a_k\left(\eta^2,\gamma^2\right) \cdot \lambda_k\left(\eta^2,\gamma^2\right)\right).\end{aligned}$$

We can numerically optimize to obtain the estimates

$$(\hat{\eta}^2_{\mathrm{ure}}, \hat{\gamma}^2_{\mathrm{ure}}) = \max_{\eta^2,\gamma^2 \geq 0} \mathrm{URE}(\eta^2,\gamma^2).$$

# EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (Hoff and Yu, 2019)

# EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (Hoff and Yu, 2019)
- Frequentist intervals shorter than standard CIs about $\hat{\tau}_r$ are impossible order-wise and difficult to obtain in practice (Chen et al., 2021).

# EB Coverage

- Valid confidence interval construction for shrinkage estimators is an open area of research (Hoff and Yu, 2019)
- Frequentist intervals shorter than standard CIs about $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ are impossible order-wise and difficult to obtain in practice (Chen et al., 2021).
- **EB coverage** is a frequently-used weaker condition
  - Implies **average coverage**: under fixed $\boldsymbol{\tau}$, a $1-\alpha$ fraction of effects are covered with high probability in large samples
  - However, some outlying effects may not be covered with $1-\alpha$ probability across repeated samples of the data

# Inference

- Advantage of hierarchical model: straightforward to extend the results of Armstrong et al. (2020) (for Stein-like shrinkers)
- Intervals have Empirical Bayes coverage guarantee *without* enforcing parametric assumptions on distribution of $\tau$ and $\xi$

# Inference

- Advantage of hierarchical model: straightforward to extend the results of Armstrong et al. (2020) (for Stein-like shrinkers)
- Intervals have Empirical Bayes coverage guarantee *without* enforcing parametric assumptions on distribution of $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$

**Definition (Robust EB Confidence Intervals (EBCIs))**

The robust EBCI for $\psi_k$, the causal effect estimate obtained from any version of double-shrinkage estimators, is

$$\psi_k \pm cva(c_k)\hat{a}_k \sqrt{\left(\hat{\lambda}_k^2 \sigma_{rk}^2 + (1-\hat{\lambda}_k)^2 \sigma_{ok}^2\right)},$$

where $\hat{a}_k$ and $\hat{\lambda}_k$ are the shrinkage factors, and $cva(c_k)$ is an inflation factor whose form is given in Armstrong et al. (2020).

# Outline

1. Assumptions and Loss Function
2. Inference
   - Positing Shrinkage Structure
   - Using a Hierarchical Model
3. Application to the WHI
4. Experimental Design

# WHI Overview

**Dataset Overview**

- Study of postmenopausal women initiated in 1991
- RCT of hormone therapy (HT) w/ 16k enrollees
- ODB w/ 50k comparable enrollees

Consider the effect of HT on coronary heart disease (CHD)

# Results

| **Subgroup Variable(s)** | **# of Strata** | Loss as a % of $\hat{\tau}_r$ Loss | | | | |
|---|---|---|---|---|---|---|
| | | $\kappa_{1+}$ | $\kappa_{2+}$ | $\hat{\psi}_{\mathbf{mm}}$ | $\hat{\psi}_{\mathbf{mle}}$ | $\hat{\psi}_{\mathbf{ure}}$ |
| CVD | 2 | 36% | 36% | 21% | <u>16%</u> | 32% |
| Age | 3 | 37% | 30% | 21% | <u>16%</u> | 34% |
| Sun | 5 | 28% | 22% | 11% | <u>9%</u> | 15% |
| CVD, Age | 6 | 39% | 42% | 21% | <u>21%</u> | 27% |
| CVD, Sun | 10 | 34% | 36% | 17% | <u>17%</u> | 19% |
| Age, Sun | 15 | 22% | 21% | 8% | <u>8%</u> | 10% |
| CVD, Age, Sun | 30 | 51% | 51% | 20% | <u>20%</u> | 20% |

Table 1: Simulation results for each stratification scheme, with an RCT sample size of $1,000$. Best-performing estimator is underlined.

# Outline

# A New Setting: Design

Can these insights inform the design of a **prospective** RCT?

- Observational study already completed, $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ obtained.
- Designing a prospective RCT of $n_r$ units
- Want to use a shrinker to combine $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ with $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. Design experiment to better complement ODB

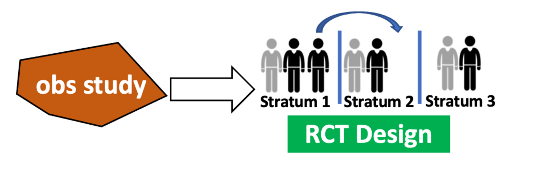
# A New Setting: Design

Can these insights inform the design of a **prospective** RCT?

- Observational study already completed, $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ obtained.
- Designing a prospective RCT of $n_r$ units
- Want to use a shrinker to combine $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ with $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. Design experiment to better complement ODB

**Goal:** choose an RCT allocation of treated and control counts per stratum, $\boldsymbol{d} = \{(n_{rkt}, n_{rkc})\}_{k=1}^{K}$, s.t. $\sum_k n_{rkt} + n_{rkc} = n_r$:

- implies how to *recruit* ...
- and *assign* treatment



# Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

# Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

# Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi)$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$

# Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \xi)$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}_{kt}^2, \hat{\sigma}_{kc}^2)\}_{k=1}^K$

# Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K$
- bias vector $\boldsymbol{\xi}$.

## Estimator and Risk

We proceed with our estimator $\boldsymbol{\kappa}_{2+}$ from the prior section:

$$\boldsymbol{\kappa}_{2+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left\{ \frac{\mathrm{Tr}(\Sigma_r^2 \boldsymbol{W})\Sigma_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} \Sigma_r^2 (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right\}_{[0,1]} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

Optimize experimental design over $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$, the risk of $\boldsymbol{\kappa}_{2+}$ under fixed $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, with

- design $\boldsymbol{d}$
- stratum potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}_{kt}^2, \hat{\sigma}_{kc}^2)\}_{k=1}^K$
- bias vector $\boldsymbol{\xi}$.

Can compute this efficiently via numerical integration (Bao and Kan, 2013), as long as $\boldsymbol{V}$ and $\boldsymbol{\xi}$ are known.

# Design Heuristics

Can estimate $\hat{\boldsymbol{V}}$ using pilot estimates obtained from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\mathrm{var}}\left(Y(1) \mid S = k\right) \qquad \text{and} \qquad \hat{\sigma}^2_{kc} = \widehat{\mathrm{var}}\left(Y(0) \mid S = k\right) .$$

# Design Heuristics

Can estimate $\hat{\boldsymbol{V}}$ using pilot estimates obtained from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\text{var}}\,(Y(1) \mid S = k) \qquad \text{and} \qquad \hat{\sigma}^2_{kc} = \widehat{\text{var}}\,(Y(0) \mid S = k)\ .$$

Design heuristics:

1. **Naïve Optimization**: Assume $\xi = 0$ and minimize $\mathcal{R}_2(\boldsymbol{d}, \hat{\boldsymbol{V}}, \xi = 0)$ over $\boldsymbol{d}$, via **greedy swap algorithm**.
2. **Robust Optimization**: Under model of Tan (2006) and a user-chosen value of sensitivity $\Gamma \geq 1$, optimize the design $\boldsymbol{d}$ under worst-case bias

# Acknowledgments

Thank you to my collaborators on this work:

- Guillaume Basse
- Mike Baiocchi
- Art Owen
- Francesca Dominici
- Luke Miratrix

Posited shrinkage structure paper available in Biometrics
Hierarchical model paper (as of Wednesday!) at arXiv:2204.06687
Design paper available at arXiv:2204.06687

# Thanks!

# References (I)

Armstrong, T. B., Kolesár, M., and Plagborg-Moller, M. (2020). Robust empirical bayes confidence intervals. *arXiv preprint arXiv:2004.03448*.

Bao, Y. and Kan, R. (2013). On the moments of ratios of quadratic forms in normal random variables. *Journal of Multivariate Analysis*, 117:229–245.

Chen, S., Zhang, B., and Ye, T. (2021). Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint arXiv:2109.10522*.

Green, E. J. and Strawderman, W. E. (1991). A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006.

Green, E. J., Strawderman, W. E., Amateis, R. L., and Reams, G. A. (2005). Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1):1–6.

Hoff, P. and Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13:94–119.

Li, K.-C. et al. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4):1352–1377.

Roehm, E. (2015). A reappraisal of Women's Health Initiative estrogen-alone trial: long-term outcomes in women 50–59 years of age. *Obstetrics and Gynecology International*, 2015.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University Stanford United States.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Xie, X., Kou, S., and Brown, L. D. (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

# Appendices

# Practical Considerations

- **Variance estimation:** In practice, $\Sigma_r$ not known. Must be estimated from data.
- **Propensity score adjustment**
  - No unconfoundedness $\implies$ propensity score adjustment can't remove all bias
  - If ODB is large, adjusting will typically be good practice. We suggest stabilized IPTW adjustments.
- **Sensitivity analysis**
  - Marginal sensitivity model of Tan (2006) summarizes degree of unmeasured confounding by a single value, $\Gamma \geq 1$
  - Can "reverse engineer" implied confounding value $\Gamma_{\text{imp}}$ when using a shrinker, via work of Zhao et al. (2019)
  - Evaluate $\Gamma_{\text{imp}}$ to obtain a ✓ or ✗ for using shrinker

# A Note on $\lambda_1^{\text{URE}}$

The true risk-minimizing shrinkage weight is given by

$$\lambda_{\text{opt}} = \frac{\text{Tr}(\Sigma_r \boldsymbol{W})}{\text{Tr}(\Sigma_r \boldsymbol{W}) + \text{Tr}(\Sigma_o \boldsymbol{W}) + \underbrace{\xi^{\text{T}} \boldsymbol{W}^2 \xi}_{\text{Not estimable from data}}},$$

but observe that

$$E\left((\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\text{T}} \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})\right) = \text{Tr}(\Sigma_r \boldsymbol{W}) + \text{Tr}(\Sigma_o \boldsymbol{W}) + \xi^{\text{T}} \boldsymbol{W}^2 \xi\,.$$

$\lambda_1^{\text{URE}}$ substitutes the quadratic form for its expectation,

$$\lambda_1^{\text{URE}} = \frac{\text{Tr}(\Sigma_r \boldsymbol{W})}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\text{T}} \boldsymbol{W} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})}\,.$$

# Guardrails

Simplicity of Algorithm 4 makes it easy to impose guardrails $\Longrightarrow$ for any invalid design, just set objective value to $\infty$.

Recommend simple guardrails for designs:

1. **Sample size**: to retain CLT, enforce

$$\min_k n_{rkt} \geq SS_{\min}, \quad \min_k n_{rkc} \geq SS_{\min}$$

2. **Detachability**: for default design $\tilde{\boldsymbol{d}} = \{\tilde{n}_{rkt}, \tilde{n}_{rkc}\}_k$ and tolerance parameter $\delta_d \geq 1$, enforce

$$\sum_k \frac{\hat{\sigma}_{kt}^2}{n'_{rkt}} + \frac{\hat{\sigma}_{kc}^2}{n'_{rkc}} \geq \delta_d \sum_k \frac{\hat{\sigma}_{kt}^2}{\tilde{n}_{rkt}} + \frac{\hat{\sigma}_{kc}^2}{\tilde{n}_{rkc}},$$

for any proposed design $\boldsymbol{d'} = \{n'_{rkt}, n'_{rkc}\}_k$.

3. **Risk reduction**: for proposed $\boldsymbol{d'} = \{n'_{rkt}, n'_{rkc}\}_k$, enforce

$$4 \max_k \left( \frac{\hat{\sigma}_{kt}^2}{n'_{rkt}} + \frac{\hat{\sigma}_{kc}^2}{n'_{rkc}} \right)^2 > \sum_k \left( \frac{\hat{\sigma}_{kt}^2}{n'_{rkt}} + \frac{\hat{\sigma}_{kc}^2}{n'_{rkc}} \right)^2.$$

# Application to the WHI

- Split RCT data into "gold" and "silver" subsets
- Gold dataset: used to obtain "gold standard" estimates of stratum treatment effects
- Repeat $1,000$ times:
  - Draw bootstrap samples:
    - $1,000$ RCT units (from silver data)
    - Observational sample (50K units)
  - Compute $L_2$ loss for $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_{1+}, \boldsymbol{\kappa}_{2+}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2$.
- Average loss over draws

# Stratification Variables

Stratify on two variables from WHI protocol (Roehm, 2015):
**Age** + **CVD** (history of cardiovascular disease)

Also include a variable unassociated with potential outcomes:
**Langley** (solar irradiance)

# Results

| Subgroup Variable(s) | # of Strata | Avg. $\hat{\tau}_r$ Loss | Loss as % of RCT-Only Loss | | | |
|---|---|---|---|---|---|---|
| | | | $\kappa_{1+}$ | $\kappa_{2+}$ | $\delta_1$ | $\delta_2$ |
| **Age** | 3 | 0.00064 | 40.1% | **34.3%** | 63.3% | 74.8% |
| **Cardiovascular disease (CVD)** | 2 | 0.00149 | 40.6% | **39.6%** | 100% | 100% |
| **Solar** | 5 | 0.00094 | 29.1% | **18.2%** | 43.1% | 52.9% |
| **Age, CVD** | 6 | 0.00574 | 25.0% | **14.0%** | 30.6% | 85.6% |
| **CVD, Solar** | 10 | 0.00803 | **20.9%** | 21.2% | 21.0% | 88.4% |
| **Age, Solar** | 15 | 0.00398 | 31.2% | 30.4% | **28.4%** | 58.4% |
| **Age, CVD, Solar** | 30 | 0.02901 | 15.8% | 16.1% | **15.7%** | 88.3% |

Table 2: Empirical risk using bootstrap samples of size 1,000 from RCT data.

# Simulations Set-Up (I)

- ODB has 20K units ($j \in \mathcal{O}$). RCT has 1,000 ($i \in \mathcal{E}$)
- Untreated potential outcomes $Y_\ell \in \{0,1\}$ for $\ell \in \mathcal{O} \cup \mathcal{E}$ sampled as indep. Bernoullis with

$$\Pr(Y_\ell(0) = 1 \mid \boldsymbol{x}_\ell) = \frac{1}{1 + e^{-\alpha - \beta^\mathsf{T} \boldsymbol{x}_\ell + \varepsilon_\ell}}, \quad \text{for } \beta = (1,1,1,1,1)^\mathsf{T}$$

for covariates $X_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_5)$, $\alpha$ chosen s.t. mean is 10%.

- Treatment variables $W_j$ for $j \in \mathcal{O}$ sampled via

$$\Pr(W_j = 1 \mid \boldsymbol{x}_j) = \frac{1}{1 + e^{-\gamma^\mathsf{T} \boldsymbol{x}_j}}, \quad \text{for } \gamma = (\sqrt{2}, \sqrt{2}, \sqrt{2}, 0, 0)^\mathsf{T}.$$

# Simulations Set-Up (II)

- Treatment effects
  - Define $k = 1, \ldots, 12$ strata based on first + second covariate
  - Assign $\tau_k$, stratum CATEs, via 3 treatment effect models:

$$\tau_k = T, \quad \tau_k = -T \times \frac{k}{K}, \quad \text{and} \quad \tau_k = T \times \left(\frac{k}{K}\right)^2$$

  - $T$ chosen so that Cohen's D in ODB equals 0.5
- Simulation structure
  - Sample ODB data a single time. Correct via SIPW.
  - Compute RCT designs under different heuristics
  - Resample RCT units $5,000$ times. For each sample, compute $L_2$ error in estimating $\boldsymbol{\tau}$ using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_2$, and $\boldsymbol{\kappa}_{2+}$

# Idealized Case: All Covariates Measured

| Est | Trt | Eq. | Ney. | Naïve | Max Bias, Γ Value 1.0 | 1.1 | 1.2 | 1.5 | Oracle |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\tau}_r$ | | 100% | **87%** | 91% | 100% | 96% | 94% | 94% | 96% |
| $\kappa_2$ | c | 82% | 48% | **44%** | 52% | 48% | 47% | 50% | 42% |
| $\kappa_{2+}$ | | 38% | 28% | **26%** | 26% | 26% | 26% | 28% | 23% |
| $\hat{\tau}_r$ | | 100% | **89%** | 92% | 95% | 94% | 95% | 97% | 104% |
| $\kappa_2$ | $\ell$ | 93% | 66% | 58% | 58% | **57%** | 60% | 64% | 50% |
| $\kappa_{2+}$ | | 59% | 51% | 45% | **43%** | 45% | 47% | 49% | 33% |
| $\hat{\tau}_r$ | | 100% | **86%** | 91% | 95% | 98% | 94% | 92% | 91% |
| $\kappa_2$ | q | 81% | 47% | **45%** | 52% | 52% | 50% | 48% | 41% |
| $\kappa_{2+}$ | | 37% | 29% | **27%** | 28% | 28% | 30% | 29% | 25% |

Table 3: Risk over 5,000 iterations of $\hat{\tau}_r, \kappa_2$, and $\kappa_{2+}$ in the case of no unmeasured confounding in the observational study. Risks are expressed as a percentage of the risk of $\hat{\tau}_r$ using an equally allocated experiment, for each of the three treatment effect models.

# Realistic Case: Third Covariate Missing

| Est | Trt | Eq. | Ney. | Naïve | Max Bias, Γ Value 1.0 | 1.1 | 1.2 | 1.5 | Oracle |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\tau}_r$ | | 100% | **90%** | 90% | 90% | 92% | 93% | 95% | 102% |
| $\kappa_2$ | c | 102% | 81% | 74% | **72%** | 72% | 72% | 77% | 69% |
| $\kappa_{2+}$ | | 96% | 80% | 74% | **71%** | 72% | 72% | 76% | 67% |
| $\hat{\tau}_r$ | | 100% | 93% | **93%** | 94% | 95% | 96% | 96% | 104% |
| $\kappa_2$ | $\ell$ | 102% | 85% | 77% | **75%** | 76% | 77% | 79% | 73% |
| $\kappa_{2+}$ | | 98% | 84% | 77% | **75%** | 76% | 76% | 79% | 71% |
| $\hat{\tau}_r$ | | 100% | **89%** | 90% | 93% | 92% | 91% | 96% | 96% |
| $\kappa_2$ | q | 101% | 74% | 69% | 68% | 68% | **67%** | 73% | 66% |
| $\kappa_{2+}$ | | 88% | 72% | 67% | 66% | 66% | **65%** | 71% | 63% |

Table 4: Risk over 5,000 iterations of $\hat{\tau}_r$, $\kappa_2$, and $\kappa_{2+}$ under various experimental designs, in the case of unmeasured confounding in the observational study via failure to measure the third covariate.

# 1. Neyman Allocation

Using stronger form of Assumption 3 (shared variances), we can estimate from the ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\mathrm{var}}\left(Y(1) \mid S = k\right) \qquad \text{and} \qquad \hat{\sigma}^2_{kc} = \widehat{\mathrm{var}}\left(Y(0) \mid S = k\right) .$$

Simplest design heuristic: use a Neyman allocation without a cost constraint, e.g.

$$n_{rkt} = \frac{n_r \cdot \hat{\sigma}^2_{kt}}{\sum_k \hat{\sigma}^2_{kt} + \hat{\sigma}^2_{kc}} \qquad \text{and} \qquad n_{rkc} = \frac{n_r \cdot \hat{\sigma}^2_{kc}}{\sum_k \hat{\sigma}^2_{kt} + \hat{\sigma}^2_{kc}} .$$

Optimizes over only the non-shrinkage portion of the risk, but reasonable in many practical settings.

# Improving Interpretability of $\boldsymbol{\kappa}_{1+}$

- Recall: $\lambda_1^{\mathrm{URE}}$ can be interpreted as an estimate of

$$\lambda_{\mathrm{opt}} = \frac{\mathrm{Tr}(\Sigma_r \boldsymbol{W})}{\mathrm{Tr}(\Sigma_r \boldsymbol{W}) + \mathrm{Tr}(\Sigma_o \boldsymbol{W}) + \boldsymbol{\xi}^\top \boldsymbol{W}^2 \boldsymbol{\xi}},$$

  true MSE-minimizing weight on $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ in a convex combination
- We can use this idea to improve interpretability of $\boldsymbol{\kappa}_{1+}$!
- **Key idea**: frame in context of sensitivity model of Tan (2006)

# Prior Work

- Marginal sensitivity model of Tan (2006) summarizes degree of unmeasured confounding by a single value, $\Gamma \geq 1$
  - $\Gamma$ bounds odds ratio of treatment prob. conditional on potential outcomes + covariates vs. covariates only
  - Related to the famous model of Rosenbaum (1987), but extends to the setting of inverse probability weighting
- Zhao et al. (2019) derive valid confidence intervals for causal estimates under the set of models indexed by any choice of $\Gamma$
  - Implicitly maps $\Gamma$ to a worst-case bias $\xi(\Gamma)$ and variance $\Sigma_O(\Gamma)$
  - Under some assumptions, allows us to obtain worst-case estimate of $\lambda_{\text{opt}}$ as a function of $\Gamma$, which we call $\lambda(\Gamma)$

# Relating the Models

- **Intuition**: larger $\Gamma$ (confounding parameter) $\implies$ optimal weight $\lambda_{\text{opt}}$ is smaller
- Let $\Gamma_{\text{imp}} = \sup\{\Gamma : \lambda(\Gamma) > \lambda_1^{\text{URE}}\}$
  - Largest value $\Gamma$ for which the optimal shrinkage factor $\lambda(\Gamma)$ is greater than our shrinkage parameter $\lambda_1^{\text{URE}}$.
- $\Gamma_{\text{imp}}$ can be used to evaluate level of shrinkage
  - If we believe true confounding level $\Gamma < \Gamma_{\text{imp}}$, then

    $$\lambda_1^{\text{URE}} \approx \lambda(\Gamma_{\text{imp}}) \leq \lambda_{\text{opt}} = \lambda(\Gamma)$$

    Hence the shrinkage level is conservative. ✓
  - If we believe $\Gamma > \Gamma_{\text{imp}}$, then estimator is overshrinking, relies too much on the observational estimate. X

# 1. Naïve Optimization Assuming $\xi = 0$ (I)

Using stronger Assumption 3 (shared var), can estimate from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\mathrm{var}}\left(Y(1) \mid S = k\right) \qquad \text{and} \qquad \hat{\sigma}^2_{kc} = \widehat{\mathrm{var}}\left(Y(0) \mid S = k\right) .$$

Define $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi}) = \mathcal{R}(\boldsymbol{\kappa}_2)$ analyzed under design $\boldsymbol{d}$, potential outcome variances $\boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kt})\}_{k=1}^K$, and error $\boldsymbol{\xi}$.

Simple heuristic: assume $\boldsymbol{\xi} = 0$. Then solve:

$$\begin{aligned} \text{minimize} \quad & \mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi}) \\ \text{subject to} \quad & \boldsymbol{\xi} = 0, \boldsymbol{V} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K, \\ & 0 < n_{rkt}, n_{rkc}, , \quad k = 1, \ldots, K, \\ & n_r = \sum_k n_{rkt} + n_{rkc} . \end{aligned} \tag{3}$$

But $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$ is not convex in the design $\boldsymbol{d}$...

# 1. Naïve Optimization Assuming $\xi = 0$ (II)

A practical approach: **greedy algorithm**. Define $\boldsymbol{d_j}$ as design on $j^{th}$ iteration, and define

$$\mathcal{D}_j = \{\boldsymbol{d}' \mid \ \boldsymbol{d}' \text{ changes one unit across strata/treatment level from } \boldsymbol{d}_j\}.$$

Run Algorithm 4 from several values of $\boldsymbol{d}_0$ and take minimum:

$$\begin{array}{l}
\texttt{Start with design } \boldsymbol{d}_0 = \{(n^{(0)}_{rkt}, n^{(0)}_{rkc})\}_k. \\
\texttt{For iteration } j = 1, 2, \ldots: \\
\quad \texttt{For each design } \boldsymbol{d}' \texttt{ in } \mathcal{D}_{j-1}: \\
\qquad \texttt{Compute } \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0)). \\
\quad \texttt{Set } \boldsymbol{d_j} = \underset{\boldsymbol{d'} \in \mathcal{D}_{j-1}}{\operatorname{argmin}} \mathcal{R}_2(\boldsymbol{d'}, \boldsymbol{V}, 0) \\
\quad \texttt{If } \mathcal{R}_2(\boldsymbol{d_j}, \boldsymbol{V}, 0) >= \mathcal{R}_2(\boldsymbol{d_{j-1}}, \boldsymbol{V}, 0) \\
\qquad \texttt{Return } d_{j-1}.
\end{array} \tag{4}$$

# Designs

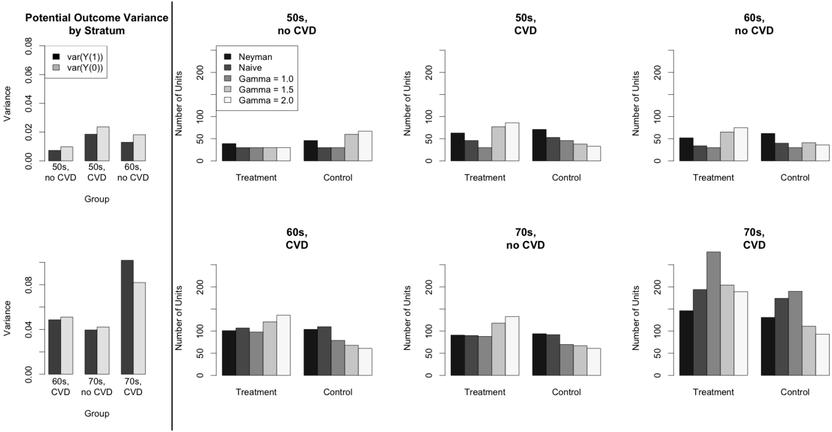



Figure 1: Allocations of $n_r = 1,000$ units in WHI with strata defined by history of CVD and age, under different design heuristics.

# Simulated Data Visualization

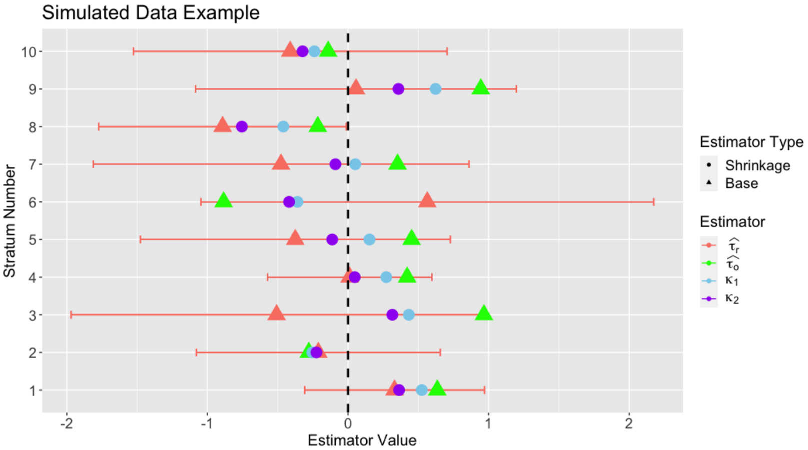



Figure 2: Simulated shrinkage between $\hat{\tau}_r$ and $\hat{\tau}_o$ with ten strata. 90% conf. sets for $\hat{\tau}_r$ in red, with $\kappa_{1+}$ and $\kappa_{2+}$ shown in circles.

## 1. Naïve Optimization Assuming $\xi = 0$

Under enhanced transportability assumption, can estimate $\hat{\boldsymbol{V}}$ using pilot estimates obtained from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\mathrm{var}}\left(Y(1) \mid S = k\right) \quad \text{and} \quad \hat{\sigma}^2_{kc} = \widehat{\mathrm{var}}\left(Y(0) \mid S = k\right) .$$

# 1. Naïve Optimization Assuming $\xi = 0$

Under enhanced transportability assumption, can estimate $\hat{\boldsymbol{V}}$ using pilot estimates obtained from ODB:

$$\hat{\sigma}^2_{kt} = \widehat{\mathrm{var}}\left(Y(1) \mid S = k\right) \qquad \text{and} \qquad \hat{\sigma}^2_{kc} = \widehat{\mathrm{var}}\left(Y(0) \mid S = k\right) .$$

Use a simple heuristic: assume $\xi = 0$.

Minimize $\mathcal{R}_2(\boldsymbol{d}, \hat{\boldsymbol{V}}, \xi = 0)$ over $\boldsymbol{d}$, via **greedy swap algorithm**.

- Swap units across strata, treatment statuses until no improvement in $\mathcal{R}_2(\boldsymbol{d}, \hat{\boldsymbol{V}}, 0)$
- Non-convexity: run from several starting points.

# 2. Heuristic Optimization Assuming Worst-Case Error Under Γ-Level Unmeasured Confounding

- Can take a more pessimistic approach using marginal sensitivity model of Tan (2006)
- For a user-chosen value of $\Gamma \geq 1$:
  - can obtain worst-case $\xi_k(\Gamma)$ using Zhao et al. (2019), and...
  - can obtain associated $\hat{\sigma}^2_{kt}$ and $\hat{\sigma}^2_{kc}$.

# 2. Heuristic Optimization Assuming Worst-Case Error Under Γ-Level Unmeasured Confounding

- Can take a more pessimistic approach using marginal sensitivity model of Tan (2006)
- For a user-chosen value of $\Gamma \geq 1$:
  - can obtain worst-case $\xi_k(\Gamma)$ using Zhao et al. (2019), and...
  - can obtain associated $\hat{\sigma}^2_{kt}$ and $\hat{\sigma}^2_{kc}$.

posit a value of $\Gamma \implies$
collect results into $\boldsymbol{V}_\Gamma$ and $\boldsymbol{\xi}_\Gamma \implies$
run greedy algorithm on $\mathcal{R}_2(\boldsymbol{d}, \boldsymbol{V}_\Gamma, \boldsymbol{\xi}_\Gamma)$ instead of $\mathcal{R}_2(\boldsymbol{d}, \hat{\boldsymbol{V}}, 0)$