Combining observational and experimental datasets using shrinkage estimators

Evan T.R. Rosenman | Guillaume Basse | Art B. Owen | Mike Baiocchi

1Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, USA
2Citadel Securities, London, UK
3Department of Statistics, Stanford University, Stanford, California, USA

Correspondence
Evan T.R. Rosenman, Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA.
Email: etrostenman@gmail.com

Abstract
We consider the problem of combining data from observational and experimental sources to draw causal conclusions. To derive combined estimators with desirable properties, we extend results from the Stein shrinkage literature. Our contributions are threefold. First, we propose a generic procedure for deriving shrinkage estimators in this setting, making use of a generalized unbiased risk estimate. Second, we develop two new estimators, prove finite sample conditions under which they have lower risk than an estimator using only experimental data, and show that each achieves a notion of asymptotic optimality. Third, we draw connections between our approach and results in sensitivity analysis, including proposing a method for evaluating the feasibility of our estimators.

KEYWORDS
causal inference, data fusion, sensitivity analysis, shrinkage

1 INTRODUCTION

The data used by researchers to assess causal effects fall broadly into two categories: experimental and observational. Well-designed experiments yield unbiased estimates of causal effects without relying on onerous statistical assumptions, making them something of a gold standard. However, experimental data are expensive to obtain and, therefore, generally involve modest sample sizes, leading to imprecise estimates of causal effects. Observational data, in contrast, are often passively collected by governments, organizations and researchers, and are therefore plentiful. However, causal effects cannot be identified from such data without making strong, unverifiable, and often unrealistic assumptions about how the intervention was assigned—or chosen by the units in the study. Sophisticated approaches can offer some degree of robustness (Lunceford & Davidian, 2004; Robins et al., 2000) but, in general, estimates of causal effects from observational studies tend to be biased. Researchers and practitioners, then, face the dilemma of choosing between estimates from experimental data, which are unbiased but imprecise, and estimates from observational studies, which are precise but biased.

This paper proposes a method for combining estimates from experiments and observational studies to improve inference. The idea of combining such data is not new and can be seen as an instance of the broader problem of data fusion (see, e.g., Bareinboim and Pearl, 2016). Kallus et al. (2018), for instance, proposes a way to leverage experimental data to reduce the confounding from an observational study by estimating a correction term and extrapolating it; while their approach does not assume ignorability in the observational data, it does make assumptions about the structure of the confounding. Lada et al. (2019) also eschew ignorability, but makes otherwise strong parametric assumptions about the potential outcomes. Finally, Athey et al. (2019) relies on additional observations to avoid dependence on ignorability.
In contrast, our approach assumes no knowledge of the confounding structure, makes no parametric assumptions, and does not require additional data. We rely on results from the Stein shrinkage literature (Green & Strawderman, 1991; Green et al., 2005; Stein, 1956) to adaptively shrink the biased observational estimate toward the unbiased experimental estimate. We prove that under certain testable conditions, the resulting estimators yield substantially lower mean squared errors than the experimental and observational estimators. We propose confidence intervals for our estimators and show that they behave well empirically.

2 | RELATED LITERATURE ON SHRINKAGE

Though they were not focused on questions of causality, Green and Strawderman (1991) addressed the question of combining biased and unbiased estimators in the empirical Bayes (EB) framework. They consider two K-dimensional multivariate normal vectors, \( \hat{\tau}_r \) and \( \hat{\tau}_o \), such that \( \hat{\tau}_r \) has mean \( \theta \) and \( \hat{\tau}_o \) has mean \( \theta - \xi \). The vectors are assumed homoscedastic with covariance matrices \( \Sigma_r = \sigma^2 I_K \) and \( \Sigma_o = \nu^2 I_k \). The goal is to estimate \( \theta \) under the \( L_2 \) loss. The authors propose the estimator

\[
\delta = \hat{\tau}_o + \left(1 - \frac{(K-2)\sigma^2}{||\hat{\tau}_o - \hat{\tau}_r||^2} \right)(\hat{\tau}_r - \hat{\tau}_o), \tag{1}
\]

and show that it dominates \( \hat{\tau}_r \) in terms of risk as long as \( K \geq 3 \). Unsurprisingly, if \( ||\xi||^2 \) is very small, the estimator underperforms a simple precision-weighted estimator. Yet, unlike the precision-weighted estimator, the proposed estimator has bounded risk as the biases grow.

The follow-up, Green et al. (2005), considers the heteroscedastic setting where \( \Sigma_r = \text{diag}(\sigma^2_{rk}) \). No assumptions are placed on \( \Sigma_o \). The authors first derive an estimator under the precision-weighted squared-error loss, in which the squared coordinate errors are scaled by the corresponding \( 1/\sigma^2_{rk} \) term. It can be shown that the estimator

\[
\delta_1 = \hat{\tau}_o + \left(1 - \frac{a}{(\hat{\tau}_r - \hat{\tau}_o)^T \Sigma_r^{-1} (\hat{\tau}_r - \hat{\tau}_o)} \right)(\hat{\tau}_r - \hat{\tau}_o), \tag{2}
\]

dominates \( \hat{\tau}_r \) under this loss, if \( \hat{\Sigma} \) is perfectly estimated, as long as \( 0 < a < 2(K-2) \).

For the more conventional-squared error loss, the authors propose a different estimator,

\[
\delta_2 = \hat{\tau}_o + \left( I_K - \frac{a \Sigma_r^{-1}}{(\hat{\tau}_r - \hat{\tau}_o)^T \Sigma_r^{-1} (\hat{\tau}_r - \hat{\tau}_o)} \right)(\hat{\tau}_r - \hat{\tau}_o). \tag{3}
\]

\( \delta_2 \) can similarly be shown to dominate \( \hat{\tau}_r \) if \( 0 < a < 2(K-2) \). The shrinkage parameter \( a \) is optimized at \( K-2 \) for \( \delta_2 \), while it depends on the value of \( \xi \) for \( \delta_1 \). Absent information about \( \xi \), however, the authors default to using \( a = K-2 \) for this estimator as well.

Chen et al. (2015) considered a similar problem: shrinkage between a regression coefficient vector estimated on a high-quality dataset and a low-quality dataset. The relevant objects for shrinkage are \( \hat{\delta}_S \), which is a normally distributed, unbiased, and homoscedastic estimate of the true coefficient vector, obtained from the high-quality dataset; and \( \hat{\delta}_B \), which is a biased estimator obtained from the low-quality dataset. The authors explore a Stein shrinkage approach. They consider an estimator, termed \( \hat{\delta}_{JS, B+} \), whose structure is identical to \( \delta \). Their main focus, however, is an approach that fits separate regressions to each dataset while penalizing the difference between model predictions from each fit. This method requires greater dimension to obtain an inadmissibility result, but performs much better in simulations when the bias is small.

3 | NOTATION, ASSUMPTIONS, AND SET-UP

3.1 | Setup

Suppose we have access to an observational study with units \( j \) in indexing set \( \mathcal{O} \) such that \( |\mathcal{O}| = n_o \). Each unit in the observational study is associated with four random quantities: a vector of observed covariates \( X_i \in \mathbb{R}^p \); a binary treatment assignment \( W_j \in \{0, 1\} \), where \( W_j = 1 \) indicates that the unit receives treatment and \( W_j = 0 \) indicates that the unit is untreated; and unseen potential outcomes \( Y_{j1}, Y_{j0} \in \mathbb{R} \), which represent the unit’s value for an outcome of interest in the presence or the absence of treatment, respectively. For each unit \( j \in \mathcal{O} \), we assume

\[
(X_j, W_j, Y_{j1}, Y_{j0}) \overset{i.i.d.}{\sim} F_O \tag{4}
\]

for some super-population distribution \( F_O \). We denote as \( E_O \) and \( \text{var}_O \) the expectation and variance operators under the distribution \( F_O \).

We define the randomized trial data analogously. That is, we have access to randomized controlled trial (RCT) units \( i \in \mathcal{R} \) and \( |\mathcal{R}| = n_r \). For \( i \in \mathcal{R} \), the quartets are drawn as

\[
(X_i, W_i, Y_{i1}, Y_{i0}) \overset{i.i.d.}{\sim} F_R \tag{5}
\]

for a distribution \( F_R \) with associated expectation and variance operators \( E_R \) and \( \text{var}_R \).
3.2 | Assumptions and loss function

We suppose that a stratification scheme, based on the observed covariates \(X_\ell, \ell \in \Theta \cup R\), is known. There are, \(k = 1, \ldots, K\) strata and each has an associated population weight \(w_1, \ldots, w_K\). Each individual \(\ell \in \Theta \cup R\) has an associated stratum indicator \(S_\ell \in \{1, \ldots, K\}\), where \(S_\ell = k \iff X_\ell \in X_k\) for some set of covariate values \(X_k\). For ease of notation, we also define indexing subsets \(\Theta_k, R_k\) (with cardinalities \(n_{o_k}, n_{r_k}\)) to identify sampled units in each stratum. We make simple assumptions about the allocation to treatment in the two studies.

**Assumption 1** (Allocations to treatment). Under \(F_O\), we have

\[
W_i Y_i(1), Y_i(0) \mid X_i,
\]

that is, unconfoundedness does not hold in the observational study. Under \(F_R\), treatment is allocated independently of the covariates.

Under Assumption 1, the selection bias in the observational study depends on variables that are not measured. Hence, we cannot use a statistical adjustment to recover an unbiased estimator of the true causal effect in the observational study. Nonetheless, we can proceed under one additional assumption.

**Assumption 2** (Common treatment effect). The conditional average treatment effect (CATE) in each stratum is identical between the two populations, that is, for all \(k = 1, \ldots, K\):

\[
\tau_k \equiv E_R(Y(1) - Y(0) \mid S = k) = E_O(Y(1) - Y(0) \mid S = k).
\]

Denote the target of estimation as \(\tau = (\tau_1, \ldots, \tau_K)\).

Assumption 2 requires that our stratification capture meaningful structure in subgroup heterogeneity that manifests in both the observational and experimental populations. These subgroups may be known a priori, or they may be discovered by deploying a modern method used for heterogeneous treatment effect estimation (Hill, 2011; Wager & Athey, 2018). The assumption’s plausibility depends on correct identification of the relevant subgroups.

We note that a closely related setting can be found in the work of Dimmery et al. (2019), which considers a multi-arm trial involving \(K\) different potential treatments. Our results can also be applied to this setting, in which the strata are defined not by heterogeneous treatment effects but by the treatments themselves: within stratum \(k\), all of the units receive treatment option \(k\) or are assigned to a control condition. In this case, the observational dataset would be an agglomeration of observational datasets in which the analogous treatments were available to units. Assumption 2 would become an assumption that each potential treatment has a treatment effect that is transportable across the observational and experimental datasets. In either case, Assumption 2 allows us to define a target of estimation in the vector of treatment effects \(\tau\).

Under Assumption 2, we consider our aggregate loss. We are interested in the individual causal effects within each stratum \(k\), rather than an overall average treatment effect. In full generality, we define our loss function as

\[
L(\tau, \tilde{\tau}) = \frac{1}{K} \sum_k d_k (\tilde{\tau}_k - \tau_k)^2 \quad \text{where} \quad d_k > 0, \sum_k d_k = 1.
\]

The stratum weights \(d_k\) reflect the relative importance placed on the accuracy of our estimator in each particular stratum. Typically, we would want \(d_k \approx w_k\), where \(w_k\) is the population weight of stratum \(k\) for a target population of interest. Lacking this, we can instead use the observational data to define a surrogate weight, \(d_k = n_{o_k}/n_o\). We denote as \(D\) the diagonal matrix whose entries are given by the \(d_k/K\), such that \(L(\tau, \tilde{\tau}) = (\tilde{\tau} - \tau)^T D (\tilde{\tau} - \tau)\).

3.3 | Estimator distributions

We define the estimators

\[
\hat{\tau}_{o_k} = \frac{\sum_{\ell \in \Theta_k} W_i Y_i}{\sum_{\ell \in \Theta_k} W_i} - \frac{\sum_{\ell \in \Theta_k} (1 - W_i) Y_i}{\sum_{\ell \in \Theta_k} (1 - W_i)} \quad \text{and}
\]

\[
\hat{\tau}_{r_k} = \frac{\sum_{\ell \in R_k} W_i Y_i}{\sum_{\ell \in R_k} W_i} - \frac{\sum_{\ell \in R_k} (1 - W_i) Y_i}{\sum_{\ell \in R_k} (1 - W_i)}.
\]

Denote \(\hat{\tau}_o = (\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK})\) and \(\hat{\tau}_r\) analogously.

We assume sufficient sample sizes and regularity conditions such that a central limit theorem holds for \(\hat{\tau}_r\). For more details on the technical conditions for this result, see, for example, Li and Ding (2017). Hence, we have approximately \(\hat{\tau}_r \sim \mathcal{N}(\tau, \Sigma_r)\).

We need not make assumptions about the distribution of \(\hat{\tau}_o\), though we denote its mean as \(\tau + \xi\), where \(\xi\) represents a \(K\)-dimensional bias parameter. The covariance matrix is denoted \(\Sigma_o\). The bias results from correlation between the potential outcomes \((Y_1(1), Y_1(0))\) and the treatment indicators within each stratum \(k\). Our assumptions imply that \(\Sigma_o\) and \(\Sigma_r\) will be diagonal matrices. We denote the diagonal entries of \(\Sigma_o\) as \(\sigma_{o1}^2, \ldots, \sigma_{oK}^2\) with analogous definitions for \(\Sigma_r\).
4 | PROPOSED ESTIMATORS

4.1 | Preliminaries

We consider how to estimate $\tau$ via shrinkage between $\hat{\tau}_r$ and $\check{\tau}_o$. We will use the term “estimator” to refer to functions of $\hat{\tau}_r$, $\check{\tau}_o$, and $\Sigma_r$. The matrix $\Sigma_r$ is not actually known to the researcher so, in practice, a plug-in estimator will be used, as will be discussed in Section 5. We first generalize a result from Strawderman (2003).

**Theorem 1** (Estimator risk). Let $Y$ be a random variable, $Z \sim \mathcal{N}(\theta, \Sigma)$, and $L(\theta, v) = (v - \theta)^\top D(v - \theta)$ where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$ and $D = 1/K \cdot \text{diag}(d_1, \ldots, d_K)$ is a diagonal weight matrix quantifying the relative importance of the $K$ components. Then for

$$\kappa(Z, Y) = Z + \Sigma g(Z, Y),$$

where $g(Z, Y)$, our “shrinkage function,” is a function of $Z$ and $Y$ that is differentiable, satisfying $E(||g||^2) < \infty$, we have

$$R(\theta, \kappa(Z, Y)) = E(L(\theta, \kappa(Z, Y))) = \frac{1}{K} \left( \text{tr}(\Sigma D) + \sum_{k=1}^K \sigma_k^4 d_k \left( g_k^2(Z, Y) + 2 \frac{\partial g_k(Z, Y)}{\partial Z_k} \right) \right)$$

(11)

**Proof.** See the Supporting information, Section 3.1.

All subsequent proofs can be found in the Supporting information, Section 3.

From Theorem 1, we can obtain a generalization of Stein’s unbiased risk estimate (SURE) (Stein, 1981) for our setting.

$$\text{URE}(\theta, \kappa(Z, Y)) = \frac{1}{K} \left( \text{tr}(\Sigma D) + \sum_{k=1}^K \sigma_k^4 d_k \times \left( g_k^2(Z, Y) + 2 \frac{\partial g_k(Z, Y)}{\partial Z_k} \right) \right).$$

(12)

We derive estimators using this unbiased risk estimate as follows: first, we will posit a structure for a proposed shrinkage estimator; second, we will derive a functional form for the shrinkage factor by optimizing the unbiased risk estimate assuming known shrinkage factors.

4.2 | $\kappa_1$, Common shrinkage factor

4.2.1 | Estimator derivation

We consider shrinkage estimators which share a common shrinkage factor across components. Denote a generic function as

$$\kappa(\lambda, \hat{\tau}_r, \check{\tau}_o) = \hat{\tau}_r - \lambda(\hat{\tau}_r - \check{\tau}_o),$$

(13)

where $\lambda$ is our common shrinkage factor.

For our second step, we will select $\lambda$ by minimizing the unbiased risk estimate. This approach has substantial precedent in the literature (see, e.g., Li et al., 1985; Xie et al., 2012). Supposing $\lambda$ is fixed ahead of time, our shrinkage function $g(\hat{\tau}_r, \check{\tau}_o)$ is simply

$$g(\hat{\tau}_r, \check{\tau}_o) = -\lambda \Sigma^{-1}(\hat{\tau}_r - \check{\tau}_o).$$

(14)

Then,

$$g_k(\hat{\tau}_r, \check{\tau}_o) = -\frac{\lambda}{\sigma^2_k}(\hat{\tau}_{rk} - \check{\tau}_{ok}) \quad \text{and} \quad \frac{\partial g_k}{\partial \hat{\tau}_{rk}} = -\frac{\lambda}{\sigma^2_k}.$$  

(15)

Plugging these results into Equation (12), the unbiased risk estimate is

$$\text{URE}(\tau, \kappa(\lambda, \hat{\tau}_r, \check{\tau}_o)) = \frac{1}{K} \left( \text{tr}(\Sigma D) + \sum_{k=1}^K \sigma_k^4 d_k \times \left( \frac{\lambda^2}{\sigma^2_k} \left( \hat{\tau}_{rk} - \check{\tau}_{ok} \right)^2 - \frac{2\lambda}{\sigma^2_k} \right) \right)$$

(16)

$$= \frac{1}{K} \left( \text{tr}(\Sigma D) + \lambda^2(\check{\tau}_o - \hat{\tau}_o)^\top D(\check{\tau}_o - \hat{\tau}_o) - 2\lambda \text{tr}(\Sigma D) \right).$$

This expression is strictly convex in $\lambda$ as long as $\hat{\tau}_r \neq \check{\tau}_o$. We seek to find

$$\lambda_1^{\text{URE}} = \min_{\lambda} \text{URE}(\tau, \kappa(\lambda, \hat{\tau}_r, \check{\tau}_o)).$$

(17)

Simple calculus tells us the unbiased risk estimate achieves its minimum at

$$\lambda_1^{\text{URE}} = \frac{\text{tr}(\Sigma D)}{(\hat{\tau}_o - \hat{\tau}_r)^\top D(\hat{\tau}_o - \hat{\tau}_r)},$$

(18)

giving us the estimator

$$\kappa_1 = \kappa(\lambda_1^{\text{URE}}, \hat{\tau}_r, \check{\tau}_o) = \hat{\tau}_r - \frac{\text{tr}(\Sigma D)}{(\hat{\tau}_o - \hat{\tau}_r)^\top D(\hat{\tau}_o - \hat{\tau}_r)}(\hat{\tau}_r - \check{\tau}_o).$$

(19)
4.2.2 | Intuition

This estimator generalizes the estimator of Green and Strawderman (1991) to the heteroscedastic, weighted-loss case. It also shares a somewhat intuitive explanation with their estimator. Suppose an oracle provided the true values of $\Sigma$, $\tau$, and $\xi$. Then, we could explicitly compute the risk, under weighted squared error loss, of $\kappa(\lambda, \hat{\tau}, \hat{\tau}_o)$ as defined in Equation (13). The risk is given by

$$R(\tau, \kappa(\lambda, \hat{\tau}, \hat{\tau}_o)) = (1 - \lambda)^2 \text{tr}(\Sigma\tau D) + \lambda^2 \text{tr}(\Sigma_o D) + \lambda^2 \xi^T D \xi.$$  

(20)

This expression has minimizer

$$\lambda_{\text{opt}} = \frac{\text{tr}(\Sigma D)}{\text{tr}(\Sigma D) + \text{tr}(\Sigma_o D) + \xi^T D \xi}.$$  

(21)

$\lambda_{\text{opt}}$ does not define a usable shrinkage estimator, because $\Sigma$, $\Sigma_o$, and $\xi$ are not known a priori. In the case of $\Sigma$, we can substitute a plug-in estimator derived from the experimental data for the true covariance matrix. However, we cannot obtain a simple plug-in estimates for $\xi$ from the observational data, because the magnitude of the bias is unknown.

However, we observe that

$$E(\hat{\tau}_o^T \Sigma \hat{\tau}_r - \hat{\tau}_o^T \Sigma \hat{\tau}_r) = \text{tr}(\Sigma D) + \text{tr}(\Sigma_o D) + \xi^T D \xi.$$  

(22)

Hence, we can obtain a reasonable estimate of the entire denominator by substituting the term $(\hat{\tau}_o - \hat{\tau}_r)^T D (\hat{\tau}_o - \hat{\tau}_r)$ for its own expectation. This is the intuition behind using $\lambda_{\text{opt}}$ as our best guess of $\lambda_{\text{opt}}$.

We consider some modifications to our estimators. We can restrict our shrinkage factor to lie between 0 and 1, an improvement also applied in Green and Strawderman (1991) and Green et al. (2005) and based on results in Baranchik (1964). Some reorganization allows us to write the estimator as

$$\kappa_{1+} = \hat{\tau}_o + \left(1 - \lambda_{1\text{URE}}\right) \hat{\lambda}(\hat{\tau}_r - \hat{\tau}_o).$$  

(23)

Another plausible modification is to optimize a scaling value $\alpha$, applied to our shrinkage factor, by again minimizing the unbiased risk estimate. As motivation, observe that the risk of $\kappa_1$ is not obtained via the expectation of Equation (16) evaluated at $\lambda_{1\text{URE}}$. This is because $\lambda_{1\text{URE}}$ is not actually known a priori, but rather it is estimated from the data; hence, we pay an additional risk penalty. Accounting for this additional penalty, it may be preferable to shrink by less than $\lambda_{1\text{URE}}$. However, in many practical cases, we find that incorporating a scaling value $\alpha$ yields muted benefits. Typically, these “corrected” estimators yield slight performance improvements in the case when $\hat{\tau}_r$ is homoscedastic or nearly-homoscedastic, and the weights $d_k$ are roughly constant; but modest performance degradations in the case when the variances $\var{\sigma}_{rk}$ or weights $d_k$ vary meaningfully across strata. For this reason, we do not recommend this approach in general. Further discussion can be found in the Supporting information, Section 4.3.

4.2.3 | Useful properties

In practice, given an experimental and observational dataset, an analyst would want to know whether to rely exclusively on the experimental data, or whether to use a shrinkage estimator. The following lemma gives us a testable condition under which $\kappa_1$ (and thus also $\kappa_{1+}$) is strictly better than $\hat{\tau}_r$ in terms of risk.

**Lemma 1.** Suppose that $4 \max_k d_k \var{\sigma}_{rk}^2 < \sum_k d_k \var{\sigma}_{rk}^2$. Then, $\kappa_1$ dominates $\hat{\tau}_r$ under our loss function.

For practical use, we substitute our estimate $\hat{\sigma}_{rk}^2$ for $\var{\sigma}_{rk}^2$ in the given condition. The condition used in Lemma 1 requires that the dimension $K$ be at least five in the homoscedastic, equal-weights case. If the variances or weights vary across strata, we may require more than five strata. While this condition allows analysts to test whether a risk reduction is guaranteed, we note that—in simulations reported in the Appendix—$\kappa_{1+}$ often outperforms $\hat{\tau}_r$ even when the condition in Lemma 1 fails to hold.

Beyond this finite sample condition, $\kappa_{1+}$ also possesses the following desirable property in the asymptotic regime.

**Theorem 2** ($\kappa_{1+}$ Asymptotic risk). Under the following (mild) conditions,

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k \var{\sigma}_{rk}^2 < \infty,$$

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k^2 \var{\sigma}_{rk}^2 < \infty,$$

and

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k^2 \var{\sigma}_{rk}^4 < \infty,$$

(24)

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

4.3 | $\kappa_2$, Variance-weighted shrinkage factors

We may instead want to choose shrinkage factors on a component-by-component basis. In the ideal case, we could choose component-wise shrinkage factors to trade
off between the inaccuracy of $\hat{\tau}_r$ and $\hat{\tau}_{ok}$. However, we cannot obtain plug-in estimators for $\xi$, making it difficult to quantify the error of each component of the vector of observational study estimates.

A more manageable heuristic is variance-weighted shrinkage on each entry in $\hat{\tau}_r$: we shrink each entry $\hat{\tau}_{rk}$ toward its counterpart $\hat{\tau}_{ok}$ by a factor proportional to its variance. The estimator relies more heavily on the randomized trial estimate for entries $k$ for which $\hat{\sigma}_{rk}^2$ is small, and more heavily on the observational estimate for entries $k$ for which $\hat{\sigma}_{rk}^2$ is large. Such an estimator forms an interesting contrast to $\delta_2$ from Green et al. (2005). $\delta_2$ exhibits the opposite behavior: it shrinks by a factor proportional to precision, such that components $\hat{\tau}_{rk}$ estimated with lower variance are shrunk more toward their counterparts $\hat{\tau}_{ok}$.

A generic, variance-weighted estimator takes the form

$$\kappa(\lambda \Sigma_r, \hat{\tau}_r, \hat{\tau}_o) = \hat{\tau}_r - \lambda \hat{\Sigma}_r (\hat{\tau}_r - \hat{\tau}_o).$$ (25)

Plugging Equation (25) into Equation (12) yields an unbiased risk estimate, which we can then minimize to obtain the optimal choice of $\lambda$ for the variance-weighted case. We obtain

$$\lambda_{2,\text{URE}} = \arg \min_\lambda \text{URE}(\tau, \kappa(\lambda \Sigma_r, \hat{\tau}_r, \hat{\tau}_o))$$

$$= \frac{\text{tr}(\Sigma_r^2 D)}{(\hat{\tau}_r - \hat{\tau}_o)^T \Sigma_r^2 D (\hat{\tau}_o - \hat{\tau}_r)},$$ (26)

yielding the estimator

$$\kappa_2 = \kappa(\lambda_{2,\text{URE}}, \hat{\tau}_r, \hat{\tau}_o) = \hat{\tau}_r - \frac{\text{tr}(\Sigma_r^2 D) \Sigma_r}{(\hat{\tau}_o - \hat{\tau}_r)^T \Sigma_r^2 D (\hat{\tau}_o - \hat{\tau}_r)} (\hat{\tau}_o - \hat{\tau}_r),$$ (27)

and its positive-part analog

$$\kappa_{2+} = \hat{\tau}_o + \left(1 - \frac{\text{tr}(\Sigma_r^2 D) \Sigma_r}{(\hat{\tau}_o - \hat{\tau}_r)^T \Sigma_r^2 D (\hat{\tau}_o - \hat{\tau}_r)}\right) (\hat{\tau}_o - \hat{\tau}_r).$$ (28)

Using the same approach as in Section 4.2, we can derive the following finite sample result for $\kappa_{2+}$.

**Lemma 2.** Suppose $4 \max_k d_k \sigma_{rk}^4 < \sum_k d_k \sigma_{rk}^2$. Then $\kappa_2$ dominates $\hat{\tau}_r$ under our loss function.

We can also obtain a similar asymptotic result.

**Theorem 3 (\kappa_2) Asymptotic risk.** Under the following (mild) conditions,

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k^2 \sigma_{rk}^6 \epsilon_k^2 < \infty,$$

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k^2 \sigma_{rk}^8 \sigma_{ok}^2 < \infty,$$

$$\limsup_{K \to \infty} \frac{1}{K} \sum_k d_k^2 \sigma_{rk}^4 < \infty,$$ (29)

in the limit $K \to \infty$, $\kappa_{2+}$ has the lowest risk among estimators using a variance-weighted shrinkage factor across components.

### 5 PRACTICAL CONSIDERATIONS

#### 5.1 Variance estimation

In practice, $\Sigma_r$ will not be known. As in Green et al. (2005), we suggest replacing it with an estimate, $\hat{\Sigma}_r$. Because we are treating the potential outcomes as a draw from a superpopulation, the standard estimator

$$\hat{\sigma}_{rk}^2 = \frac{1}{n_{rk}t} \sum_{i \in R_k} W_i (Y_i - \hat{Y}_{rki})^2 + \frac{1}{n_{rk}c} \sum_{i \in R_k} (1 - W_i) (Y_i - \hat{Y}_{rkc})^2$$ (30)

where

$$n_{rk}t = \sum_{i \in R_k} W_i,$$ and $$n_{rk}c = \sum_{i \in R_k} 1 - W_i,$$

$$\hat{Y}_{rki} = \frac{1}{n_{rk}t} \sum_{i \in R_k} W_i Y_i,$$ and $$\hat{Y}_{rkc} = \frac{1}{n_{rk}c} \sum_{i \in R_k} (1 - W_i) Y_i$$ (31)

is unbiased for $\sigma_{rk}^2$.

We note that the testable conditions expressed in Lemmas 1 and 2 are written in terms of the true variance, rather than an estimated variance. The estimated variance can, of course, serve as a proxy, but the condition may hold only approximately if the values $\hat{\sigma}_{rk}^2$ are estimated using a small number of units. It is also plausible that using an estimated variance could impact the performance of $\kappa_{1+}$ and $\kappa_{2+}$. However, in simulations, we do not see any systematic decline in performance when using $\hat{\Sigma}_r$ in place of $\Sigma_r$. We include such a performance comparison in our simulations in Section 7.

#### 5.2 Propensity score adjustment

Because treatment is not randomized in the observational data, these data will suffer from selection bias. As
discussed in Section 1, we do not assume unconfoundedness. Thus, we cannot consistently estimate the propensity score from the observational data, and so conditioning on the estimated propensity score will not remove all of the selection bias.

Using any standard adjustment method from the causal inference literature—for example, matching, stratification, or regression (Imbens & Rubin, 2015)—we can reduce, though not fully eliminate, bias. Because the observational study is assumed much larger than the randomized trial, any increase in variance from the propensity adjustment will likely be outweighed by this decrease in bias.

Estimation of the propensity score will depend on the problem set-up. If the strata \( k \) represent different treatments, then a different propensity model should be fit in each arm. If they represent subgroups with different treatment effects, then a single propensity model can be fit. There are many common tactics for adjusting causal estimates using an estimated propensity score. We advocate stabilized inverse probability weighted estimation (see, e.g., Robins et al., 2000), where

\[
\hat{\tau}_{ek} = \sum_{i \in \Omega_k} \frac{W_i Y_i}{\hat{p}_i} \left( \sum_{i \in \Omega_k} \frac{W_i}{\hat{p}_i} \right)^{-1} - \sum_{i \in \Omega_k} \frac{(1 - W_i) y_i}{1 - \hat{p}_i} \left( \sum_{i \in \Omega_k} \frac{1 - W_i}{1 - \hat{p}_i} \right)^{-1}
\]  

(32)

This is simply the Horvitz–Thompson inverse probability weighted estimator with normalized weights (Horvitz & Thompson, 1952). This method will admit a straightforward sensitivity analysis, allowing analysts to better quantify the amount of bias implied by the shrinkage estimator.

5.3 Constraining confidence intervals

The construction of valid confidence intervals for shrinkage estimators is an open area of research (Armstrong et al., 2022; Hansen, 2016; Hoff, 2022). The standard approach is to use a relaxed notion of interval coverage, known as “Empirical Bayes coverage” (Armstrong et al., 2022; Morris, 1983). Many shrinkage estimators are derived under an imagined hierarchical model, in which the true effects (\( \tau \), in our setting) are themselves sampled from a distribution. EB coverage requires that the \( 1 - \alpha \) coverage rate holds over repeated sampling of both the data and the true effects themselves. For any particular values of the true effects \( \tau \), some entries may be under-covered by a procedure achieving EB coverage, while other entries may be over-covered. However, if the procedure is derived under a distribution that is reasonably close to the empirical distribution of the true effects \( \tau \), this implies that the EB intervals should cover roughly a \( 1 - \alpha \) fraction of the entries of \( \tau \) (Hoff, 2022).

EB coverage is strictly weaker than standard frequentist coverage, under which we would require the \( 1 - \alpha \) coverage rate to hold over repeated sampling of the data for any values of \( \tau \). Armstrong et al. (2022)—drawing on an argument originated in Pratt (1961) and expanded upon in Armstrong and Kolesár (2018)—argue that, in most standard settings, relaxing the usual frequentist coverage criterion is necessary to obtain meaningful efficiency gains from EB methods.

If researchers are willing to accept EB coverage, we suggest constructing intervals using the bootstrap procedures introduced in Laird and Louis (1987). This work builds on the intervals developed in Morris (1983), which are frequentist approximations to Bayesian posterior credible intervals. The bootstrap procedure of Laird and Louis (1987) imposes fewer parametric assumptions about the data-generating process. Under this procedure, standard errors for the stratum-specific causal estimates take the form

\[
s_k = \left( \hat{\sigma}^2_{rk} \left( 1 - \frac{K - 1}{K} \hat{\lambda} \right) + \hat{B}^2_k \right)^{1/2}, \quad k = 1, \ldots, K,
\]

(33)

where \( \hat{\lambda} \) is equal to \( \lambda^{URE}_{1+2} \) for \( \kappa_{1+} \) and equal to the \( k \)th diagonal entry of \( \lambda^{URE}_{2+2} \) for \( \kappa_{2+} \). \( \hat{B}^2_k \) is the empirical variance, across bootstrap replicates, of the \( k \)th entry of the estimator. We use the “Type III” bootstrap suggested by Laird and Louis. Note that, as in Green et al. (2005), we modify the first term of the standard error because we are considering shrinkage between biased and unbiased estimators, rather than across entries of a multivariate normal vector. With these standard errors, \( Z \)-intervals can be constructed at any desired level of \( \alpha \) in the usual way.

If researchers require the more stringent benchmark of frequentist coverage, the standard approach is to recenter confidence intervals derived from frequentist procedures at the shrunk point estimates (Casella et al., 2012; Hoff, 2022; Stein, 1962). This approach works in general when shrinking across multiple unbiased parameter estimates (as in the setting of the James–Stein estimator). However, when shrinking between biased and unbiased estimators, it is easy to construct examples where it fails to achieve frequentist coverage due to over-shrinkage toward the biased estimate. We report these intervals in our applied data analysis, but caution that improved procedures for achieving frequentist coverage are an important area for further research.
6 | INTERPRETABLE DIAGNOSTICS

We saw in Section 4.2 that our estimator $\kappa_1$ can be interpreted as the result of plugging an estimate of the optimal shrinkage factor $\lambda_{opt}$ into Equation (13). Analogously, $\kappa_{1+}$ can be interpreted as the result of plugging $\hat{\lambda}_{opt} \equiv (1 - (1 - \lambda_{opt}^\text{URE}))$ into Equation (13). Shrinking an unbiased estimator toward a biased estimator is not commonly done in the causal inference literature; one reason, we believe, is the uninterpretability of the estimate $\hat{\lambda}_{opt}$, relative to other tools in the causal inference arsenal. To remedy this, we connect our shrinkage estimate to more familiar quantities in the causal literature. We sketch the argument in this section, and refer the reader to the Supporting information for the details.

One way to make the estimate $\hat{\lambda}_{opt}$ more interpretable is to frame it in the context of the well-known marginal sensitivity model of Tan (2006), wherein the degree of confounding is summarized by a single value, $\Gamma \geq 1$, which bounds the odds ratio of the treatment probability conditional on the potential outcomes and covariates and the treatment probability conditional only on covariates. For a given choice of $\Gamma$, recent works in the literature (Dorn & Guo, 2022; Zhao et al., 2019) have derived bounds for the worst-case bias $\xi(\Gamma)$; which, under some assumptions on the variance $\Sigma$, yields an upper bound of the denominator of Equation (21), and therefore a lower bound on the optimal shrinkage parameter, $\lambda(\Gamma)$. This gives us an interpretable connection between the optimal shrinkage and the $\Gamma$ confounding parameter in the context of the marginal sensitivity model: the more confounding is allowed by the model—that is, the larger $\Gamma$ becomes—the less we should shrink toward the observational estimate.

Now to connect this to the estimated shrinkage $\hat{\lambda}_{opt}$, let $\Gamma_{imp} = \sup\{\Gamma : \lambda(\Gamma) > \hat{\lambda}_{opt}\}$ the largest value $\Gamma$ for which the optimal shrinkage factor $\lambda(\Gamma)$ is greater than our estimate $\hat{\lambda}_{opt}$. The quantity $\Gamma_{imp}$ can serve as an interpretable diagnostic. Indeed, suppose the analyst believes that the true amount of confounding is $\Gamma < \Gamma_{imp}$. Then

$$\hat{\lambda}_{opt} \approx \lambda(\Gamma_{imp}) \leq \lambda_{opt} = \lambda(\Gamma),$$

and our estimator $\kappa_{1+}$ undershrinks toward the observational estimate $\hat{\tau}_o$. This is a conservative attitude: the analyst may proceed with our estimator. If, however, the analyst believes that the true confounding $\Gamma > \Gamma_{imp}$, then our estimator is likely to overshrink: that is, it relies too much on the observational estimate $\hat{\tau}_o$. In sum, a large value of $\Gamma_{imp}$ is desirable, since it implies that our estimator is likely to be conservatively undershrinking. In addition, since we rely on the widely used marginal sensitivity model, the value of $\Gamma_{imp}$ is directly comparable with other values of $\Gamma$ typically found in the literature, making it an intuitively interpretable quantity.

7 | SIMULATIONS USING DATA FROM THE WOMEN’S HEALTH INITIATIVE

7.1 | Setup

To evaluate our methods in practice, we conduct simulations using data from the Women’s Health Initiative (WHI), a 1991 study of the effects of hormone therapy on postmenopausal women. The study included both a randomized controlled trial and an observational study. A total of 16,608 women were included in the trial, with half randomly selected to take 625 mg of estrogen and 2.5 mg of progestin, and the remainder comprising the control. A corresponding 53,054 women in the observational component of the WHI were deemed clinically comparable to women in the trial. About a third of these women were using estrogen plus progestin, while the remaining women in the observational study were not using hormone therapy (Prentice et al., 2005).

We investigate the effect of the treatment on incidence of coronary heart disease. To evaluate our methods, we draw 1,000 bootstrap samples from the RCT component and observational component of the WHI, and compute our shrinkage estimators on each bootstrap sample. The bootstrap serves as a proxy for sampling from a super-population. Causal estimates computed on the RCT bootstrap samples are unbiased and normally distributed about the “true” causal estimates computed for the entire RCT sample. Hence, the assumptions for using our estimators are met. Moreover, we can empirically evaluate the risk of our estimators by computing their average mean squared error in estimating these true causal quantities.

To choose our subgroups for stratification, we utilize the clinical expertise of researchers in the study’s writing group. The trial protocol highlights age as an important subgroup variable to consider (Writing Group for the WHI Investigators, 1998), while subsequent work considered a patient’s history of cardiovascular disease (Roehm, 2015). We also consider Langley scatter, a measure of solar irradiance at each woman’s enrollment center, which is not plausibly related to baseline incidence or treatment effect.

Langley scatter exhibits no association with the outcome plausibly related to baseline incidence or treatment effect. Langley scatter exhibits no association with the outcome. The bootstrap serves as a proxy for sampling from a super-population. Causal estimates computed on the RCT bootstrap samples are unbiased and normally distributed about the “true” causal estimates computed for the entire RCT sample. Hence, the assumptions for using our estimators are met. Moreover, we can empirically evaluate the risk of our estimators by computing their average mean squared error in estimating these true causal quantities.

The age variable has three levels, corresponding to...
whether a woman was in her fifties, sixties, or seventies. The cardiovascular disease history variable is binary. The Langley scatter variable has five levels, corresponding to strata between 300 and 500 Langley units of irradiance. We provide brief summaries of these variables in Tables 13–15 in Section 6.3 of the Supporting information.

### 7.2 Results

We consider two settings. In the first case, we draw bootstrap samples of size 1,000 units from the RCT, while bootstrapping the entire observational study, which is approximately 50 times larger. We also adjust the observational study using a propensity score adjustment. Details on the computation of the propensity score can be found in Rosenman et al. (2022).

This setting reflects a realistic scenario, in which the analyst has access to a very small, high-quality RCT, along with a much larger observational database whose estimates are reasonably good but have some residual bias. Within each of the 1,000 bootstrap replicates, we compute the average squared error loss of our estimators in recovering the “true” causal effects within each stratum. The “true” effects are computed by applying the difference-in-means estimator within each stratum in the full RCT sample.

In Table 1, we show the results of the simulations. Each row corresponds to a different stratification scheme, generated by stratifying on a subset of our stratification variables. In the second column, we give the number of strata. The following six columns give the average mean squared error of different estimators, expressed as a percentage of the average MSE of \( \hat{\tau}_r \). Any value less than 100% indicates a loss reduction. We considered our estimators, \( \kappa_{1+} \) and \( \kappa_{2+} \), as well as \( \delta_1 \) and \( \delta_2 \), the estimators introduced in Green et al. (2005). For reference we also consider \( \hat{\tau}_o \), the observational study estimator, and \( \hat{\tau}_w \), a precision-weighted convex combination of \( \hat{\tau}_o \) and \( \hat{\tau}_r \). In this regime—in which the RCT is significantly smaller than the observational study—both of these estimators have lower MSE than \( \hat{\tau}_r \), and also outperform the shrinkers. In practice, analysts would be hesitant to use either \( \hat{\tau}_o \) or \( \hat{\tau}_w \) because these estimators are not structured to account for bias due to unmeasured confounding.

Among the shrinkers, every shrinkage estimator performs at least as well as \( \hat{\tau}_r \). In the case of just two strata (the first row), there is no shrinkage for \( \delta_1 \) and \( \delta_2 \), so each of these estimators simplifies to \( \hat{\tau}_r \); in every case, where shrinkage occurs, there is a risk improvement of at least 20%. Our estimators perform best under the first four stratification schemes. \( \kappa_{2+} \) is the top performer when there are relatively few strata (2, 3, or 5), while \( \kappa_{1+} \) does best when there are 6 strata. As the number of strata increases, the performances of \( \kappa_{1+} \), \( \kappa_{2+} \), and \( \delta_1 \) converge to similar values. The best estimator in the 10-strata condition is \( \delta_1 \), while \( \kappa_{1+} \) performs best in the 15-strata condition. In the 30-strata condition \( \kappa_{2+} \) performs best, but its improvement over \( \delta_1 \) and \( \kappa_{2+} \) is quite marginal.

We also consider a second setting in which the RCT sample size is significantly larger: 8,000 units. We again draw samples from both datasets via the bootstrap, and use a propensity score adjustment in the observational study. Results from these simulations can be found in Table 2.

Because of the larger sample size in the RCT, the observational study estimator now has larger MSE than the RCT estimator across all stratifications with three or more strata. The precision-weighted estimator also no longer uniformly outperforms the shrinkers. In this regime, \( \kappa_{2+} \) does best under the first stratification, and \( \kappa_{1+} \) under the second. In the remaining five stratifications, \( \delta_1 \) outperforms, albeit modestly in many of the stratifications. We observe that \( \kappa_{1+} \) also tends to do well as the number of strata grows. However, \( \kappa_{2+} \) does not uniformly outperform \( \hat{\tau}_r \) in this regime (in which the condition of Lemma 2 is generally not met under any of the stratifications), an important cautionary note for the use of this estimator.

Further data from these simulations can be found in the Supporting information, Section 1. In Section 1.1, we decompose the MSE discussed in Tables 1 and 2 into its

---

**TABLE 1** Simulation results for each stratification scheme, with an RCT sample size of 1,000

| Subgroup Var(s) | Number of Strata | Loss as % of \( \hat{\tau}_r \) Loss |
|-----------------|-----------------|-------------------------------------|
|                 | \( \tau_o \) | \( \hat{\tau}_w \) | \( \kappa_{1+} \) | \( \kappa_{2+} \) | \( \delta_1 \) | \( \delta_2 \) |
| CVD             | 2               | 1.05% | 9.7%  | 37.6%  | 36.9%  | 100.0% | 100.0% |
| Age             | 3               | 18.4% | 16.8% | 37.3%  | 30.1%  | 61.5%  | 72.8%  |
| Langley         | 5               | 23.7% | 20.9% | 29.4%  | 23.5%  | 40.0%  | 52.2%  |
| CVD, Age        | 6               | 36.9% | 31.5% | 38.0%  | 38.2%  | 38.3%  | 82.4%  |
| CVD, Langley    | 10              | 30.9% | 28.2% | 30.6%  | 32.5%  | 30.0%  | 87.2%  |
| Age, Langley    | 15              | 23.0% | 19.9% | 22.4%  | 23.0%  | 22.5%  | 43.1%  |
| CVD, Age, Langley | 30             | 50.3% | 43.5% | 50.3%  | 50.3%  | 50.3%  | 78.4%  |

Note: The best-performing shrinker is underlined for each stratification scheme.
TABLE 2  Simulation results for each stratification scheme, with an RCT sample size of 8,000

| Subgroup Var(s) | Number of Strata | Loss as % of $\tau_\omega$ Loss |
|-----------------|------------------|-------------------------------|
|                 |                  | $\hat{\tau}_o$ | $\hat{\tau}_w$ | $\kappa_{1+}$ | $\kappa_{2+}$ | $\delta_1$ | $\delta_2$ |
| CVD             | 2                | 73.3% | 45.5% | 72.6% | 65.5% | 100.0% | 100.0% |
| Age             | 3                | 156.0% | 74.9% | 83.7% | 93.3% | 85.0% | 91.1% |
| Langley         | 5                | 171.3% | 89.4% | 82.1% | 141.7% | 79.0% | 85.0% |
| CVD, Age        | 6                | 240.5% | 102.5% | 86.2% | 130.6% | 79.2% | 93.0% |
| CVD, Langley    | 10               | 148.4% | 67.5% | 69.9% | 106.6% | 64.1% | 95.2% |
| Age, Langley    | 15               | 122.6% | 66.8% | 63.3% | 108.3% | 60.8% | 81.2% |
| CVD, Age, Langley | 30             | 166.0% | 70.5% | 81.6% | 155.6% | 69.2% | 95.4% |

Note: The best-performing shrinker is underlined for each stratification scheme.

We find that the shrinkers behave as expected: they incorporate significant bias by shrinking toward $\hat{\tau}_o$, but, in doing so, typically see a massive reduction in variance. This generally leads to lower MSE, even if the unbiasedness property of $\hat{\tau}_w$ is sacrificed. In Section 1.2, we empirically evaluate whether the use of an estimated variance $\hat{\Sigma}_r$ (rather than the true variance $\Sigma_r$) materially impedes performance. Analogous results to those in Tables 1 and 2 can be found in the Supporting information, Tables 5 and 6, except the simulations have been conducted using shrinkers evaluated at the true values of $\Sigma_r$ rather than at $\hat{\Sigma}_r$ estimated in every bootstrap replicate. In comparing the tables, we see there is no systematic loss of performance from using $\hat{\Sigma}_r$. Performances are very nearly identical using real versus estimated RCT variances, and any deviations in performance seem as likely to increase mean squared error as to reduce it. This provides strong evidence that using estimated variances does not meaningfully hinder the performance of our estimators.

Lastly, we consider the question of inference. We again draw 1,000 bootstrap replicates from the data and compute our shrinkage estimators, along with confidence intervals using the centering and bootstrap credible interval methods described in Section 5.3. For the bootstrap, 100 replicates are generated per iteration to compute $\hat{B}_k^2$ for $k = 1, ..., K$. Within each stratum, we compute the frequency with which a 90% confidence interval covers the true effects estimated from the full RCT dataset. Coverage rates are then aggregated per stratum across the draws. We show the average and minimum coverage rates across strata, and the average interval length, for both confidence interval methods. Results for the case with 1,000 RCT units can be found in Table 3, and for the case with 8,000 units in Table 4.

At the smaller RCT sample size (Table 3), we see the expected behavior for both types of intervals. The centering intervals consistently achieve the desired frequentist coverage rate, while the bootstrap intervals consistently achieve average (i.e., EB) coverage. As expected, the bootstrap intervals undercover causal estimates from some strata and overcover others. While they are slightly longer than the centering intervals when the number of strata is small, they are consistently shorter than the centering intervals as the number of strata grows. At the larger RCT sample size (Table 4), we do not see the desired behavior for the centering intervals: their minimum coverage rates across strata are consistently below the desired nominal rate of 90%. In fact, the centering intervals even fail to achieve average coverage in a few cases. However, the bootstrap intervals continue to achieve EB coverage at the larger sample size. The bootstrap intervals are slightly longer than the centering intervals in most cases, though they are shorter under the finest stratification, in the final row. Taken together, these results suggest that EB coverage is achievable, while frequentist coverage may fail in certain data regimes.

8  FUTURE WORK

There are numerous potential extensions to this work. We have explored two shrinkage structures in this text—shrinkage by a constant factor, and shrinkage by a variance-weighted factor—but our procedure is general and can be used to derive alternative estimators. For example, we may sometimes consider estimates of $\Sigma_o$ to be reliable, despite unmeasured confounding in the observational study. If so, we could posit the structure for a shrinker that incorporates the estimated covariance matrix $\hat{\Sigma}_o$, and derive its exact form via our URE-minimization approach. Other forms of auxiliary information could similarly be utilized in tandem with URE minimization to design improved shrinkage estimators.

A related extension to this work is the development of pragmatic guidelines for choosing among shrinkers. In our WHI simulations, we saw that there was no consistent best performer among our estimators $\kappa_{1+}$ and $\kappa_{2+}$ as well as $\delta_1$ and $\delta_2$ from Green et al. (2005). Under different stratifications and different RCT sizes, both $\kappa_{1+}$ and $\delta_1$...
TABLE 3  Coverage results for estimators $\kappa_{1+}$ and $\kappa_{2+}$ when the RCT sample size is 1,000

| Subgroup Var(s) | Number of strata | Est. | Recentering coverage Mean | Min | Len | Bootstrap coverage Mean | Min | Len |
|-----------------|------------------|------|---------------------------|-----|-----|--------------------------|-----|-----|
|                 |                  |      |                           |     |     |                          |     |     |
| CVD             | 2                | $\kappa_{1+}$ | 98% | 97% | 0.101 | 100% | 100% | 0.128 |
|                 |                  | $\kappa_{2+}$ | 95% | 93% | 0.101 | 100% | 100% | 0.111 |
| Age             | 3                | $\kappa_{1+}$ | 99% | 96% | 0.071 | 100% | 100% | 0.080 |
|                 |                  | $\kappa_{2+}$ | 99% | 98% | 0.071 | 100% | 100% | 0.075 |
| Langley         | 5                | $\kappa_{1+}$ | 99% | 98% | 0.089 | 99% | 94% | 0.090 |
|                 |                  | $\kappa_{2+}$ | 100% | 99% | 0.089 | 97% | 87% | 0.080 |
| CVD, Age        | 6                | $\kappa_{1+}$ | 100% | 99% | 0.184 | 99% | 97% | 0.169 |
|                 |                  | $\kappa_{2+}$ | 98% | 94% | 0.184 | 96% | 84% | 0.130 |
| CVD, Langley    | 10               | $\kappa_{1+}$ | 100% | 99% | 0.276 | 99% | 98% | 0.202 |
|                 |                  | $\kappa_{2+}$ | 99% | 95% | 0.276 | 96% | 81% | 0.137 |
| Age, Langley    | 15               | $\kappa_{1+}$ | 100% | 99% | 0.179 | 98% | 89% | 0.120 |
|                 |                  | $\kappa_{2+}$ | 100% | 99% | 0.179 | 91% | 64% | 0.086 |
| CVD, Age, Langley | 30            | $\kappa_{1+}$ | 100% | 98% | 0.624 | 94% | 65% | 0.526 |
|                 |                  | $\kappa_{2+}$ | 100% | 98% | 0.624 | 87% | 5% | 0.157 |

Note: Intervals are computed using the recentering and bootstrap approximate credible interval approaches, for each stratification scheme.

TABLE 4  Coverage results for estimators $\kappa_{1+}$ and $\kappa_{2+}$ when the RCT sample size is 8,000

| Subgroup Var(s) | Number of strata | Est. | Recentering coverage Mean | Min | Len | Bootstrap coverage Mean | Min | Len |
|-----------------|------------------|------|---------------------------|-----|-----|--------------------------|-----|-----|
|                 |                  |      |                           |     |     |                          |     |     |
| CVD             | 2                | $\kappa_{1+}$ | 79% | 60% | 0.037 | 89% | 78% | 0.050 |
|                 |                  | $\kappa_{2+}$ | 89% | 81% | 0.037 | 97% | 95% | 0.045 |
| Age             | 3                | $\kappa_{1+}$ | 93% | 90% | 0.026 | 97% | 95% | 0.035 |
|                 |                  | $\kappa_{2+}$ | 91% | 86% | 0.026 | 96% | 93% | 0.034 |
| Langley         | 5                | $\kappa_{1+}$ | 93% | 76% | 0.033 | 96% | 84% | 0.042 |
|                 |                  | $\kappa_{2+}$ | 80% | 26% | 0.033 | 90% | 65% | 0.042 |
| CVD, Age        | 6                | $\kappa_{1+}$ | 94% | 86% | 0.067 | 95% | 92% | 0.096 |
|                 |                  | $\kappa_{2+}$ | 91% | 80% | 0.067 | 89% | 75% | 0.089 |
| CVD, Langley    | 10               | $\kappa_{1+}$ | 94% | 66% | 0.096 | 94% | 82% | 0.117 |
|                 |                  | $\kappa_{2+}$ | 94% | 82% | 0.096 | 93% | 62% | 0.108 |
| Age, Langley    | 15               | $\kappa_{1+}$ | 97% | 85% | 0.065 | 96% | 86% | 0.074 |
|                 |                  | $\kappa_{2+}$ | 93% | 66% | 0.065 | 91% | 65% | 0.075 |
| CVD, Age, Langley | 30            | $\kappa_{1+}$ | 97% | 72% | 0.197 | 92% | 72% | 0.174 |
|                 |                  | $\kappa_{2+}$ | 95% | 63% | 0.197 | 88% | 53% | 0.151 |

Note: Intervals are computed using the recentering and bootstrap approximate credible interval approaches, for each stratification scheme.

typically performed well. $\kappa_{2+}$ often performed best when the RCT size was small, but often performed worst when the RCT size was large. We aim to develop practical recommendations for settings in which each estimator (or alternatives derived via our URE-minimization procedure) can be expected to outperform.

Lastly, an area for further research is the construction of confidence intervals. We have proposed a bootstrap method for constructing intervals that achieve the desired level of EB coverage, though it is possible that an alternative procedure could obtain the same level of coverage with shorter average interval lengths. More importantly, as discussed in Section 5.3, we lack general-purpose methods for constructing confidence intervals under standard notions of frequentist coverage. Identifying procedures to construct frequentist intervals will be a significant priority in future work.

ACKNOWLEDGMENTS
The authors thank Dylan Small for pointing us toward Green and Strawderman (1991) and for providing early discussions about some of the ideas in this paper. Evan T.R.
Rosenman was supported by Google, and by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This work was also supported by the NSF under grants DMS-1521145, DMS-1407397, and IIS-1837931.

**DATA AVAILABILITY STATEMENT**

The data used in the simulations in Section 7 are available from the Women’s Health Initiative. Restrictions apply to the availability of these data, which were used under license for this paper. Data are available at https://www.whi.org/ with the permission of the Women’s Health Initiative.

**ORCID**

Evan T.R. Rosenman DOI https://orcid.org/0000-0001-5885-2925

**REFERENCES**

Armstrong, T.B. & Kolesár, M. (2018) Optimal inference in a class of regression models. *Econometrica*, 86(2), 655–683.

Armstrong, T.B., Kolesár, M. & Plagborg-Møller, M. (2022) Robust empirical bayes confidence intervals. *Econometrica*, 90(6), 2567–2602.

Athey, S., Chetty, R., Imbens, G.W. & Kang, H. (2019) The surrogate index: combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report. National Bureau of Economic Research.

Baranchik, A. (1964) Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report. Stanford University, Stanford, CA.

Bareinboim, E. & Pearl, J. (2016) Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), 7345–7352.

Casella, G. & Hwang, J.G. (2012) Shrinkage confidence procedures. *Statistical Science*, 27(1), 51–60.

Chen, A., Owen, A.B. & Shi, M. (2015) Data enriched linear regression. *Electronic Journal of Statistics*, 9(1), 1078–1112.

Dimmery, D., Bakshy, E. & Sekhon, J. (2019) Shrinkage estimators in online experiments. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2914–2922.

Dorn, J., & Guo, K. (2022) Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, (just-accepted), 1–28.

Green, E.J. & 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. & Reams, G.A. (2005) Improved estimation for multiple means with heterogeneous variances. *Forest Science*, 51(1), 1–6.

Hansen, B.E. (2016) Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1), 115–132.

Hill, J.L. (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.

Hoff, P. (2022) Coverage properties of empirical Bayes intervals: a discussion of “confidence intervals for nonparametric empirical Bayes analysis” by ignatiadis and wager. *Journal of the American Statistical Association*, 117(539), 1175–1178.

Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Imbens, G.W. & Rubin, D.B. (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. New York, NY, USA: Cambridge University Press.

Kallus, N., Puli, A.M. & Shalit, U. (2018) Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, pp. 10888–10897.

Lada, A., Peysakhovich, A., Aparicio, D., & Bailey, M. (2019) Observational data for heterogeneous treatment effects with application to recommender systems. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 199–213.

Laird, N.M. & Louis, T.A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399), 739–750.

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

Li, X. & Ding, P. (2017) General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520), 1759–1769.

Lunceford, J.K. & Davidian, M. (2004) Stratification and weighting in multiple regression and estimation of the mean of a multivariate normal distribution. *Statistical Science*, 19(4), 2972–2979.

Morrison, C.N. (1983) Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381), 47–55.

Patt, J.W. (1961) Length of confidence intervals. *Journal of the American Statistical Association*, 56(295), 549–567.

Prentice, R.L., Langer, R., Stefanick, M.L., Howard, B.V., Pettinger, M., Anderson, G., et al. (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women’s Health Initiative clinical trial. *American Journal of Epidemiology*, 162(5), 404–414.

Robins, J.M., Hernan, M.A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

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, 713295.

Rosenman, E.T., Owen, A.B., Baiocchi, M. & Banack, H.R. (2022) Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 41(1), 65–96.

Stein, C. (1956) *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Technical Report. Stanford University, Stanford, CA.

Stein, C.M. (1962) Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24(2), 265–285.

Stein, C.M. (1981) Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9, 1135–1151.

Strawderman, W.E. (2003) On minimax estimation of a normal mean vector for general quadratic loss. *Mathematical statistics*
Tan, Z. (2006) A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476), 1619–1637.

Wager, S. & Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.

Writing Group for the WHI Investigators., (1998) Design of the Women’s Health Initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1), 61–109.

Xie, X., Kou, S. & 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., & 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)*, 81(4), 735–761.

**SUPPORTING INFORMATION**

Web Appendices and Tables referenced in Sections 4, 6, and 7 are available with this paper at the Biometrics website on Wiley Online Library. A repository of code for the simulations in Section 7 (as well as additional simulations provided in the Supporting information) is also provided with this paper at the Biometrics website on Wiley Online library.

**How to cite this article:** Rosenman, E.T.R., Basse, G., Owen, A.B. & Baiocchi, M. (2023) Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79, 2961–2973. https://doi.org/10.1111/biom.13827