Electronic Journal of Statistics
Vol. 17 (2023) 3406–3442
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2179

Designing experiments toward shrinkage estimation

Evan T. R. Rosenman[1] **and Luke Miratrix**[2]

[1]*Department of Mathematical Sciences, Claremont McKenna College*
e-mail: erosenman@cmc.edu

[2]*Graduate School of Education, Harvard University*
e-mail: luke_miratrix@gse.harvard.edu

Abstract: How can increasingly available observational data be used to improve the design of randomized controlled trials (RCTs)? We seek to design a prospective RCT, with the intent of using an Empirical Bayes estimator to shrink the causal estimates from our trial toward causal estimates obtained from an observational study. We ask: how might we design the experiment to better complement the observational study in this setting?

We show that the risk of such shrinkage estimators can be computed efficiently via numerical integration. We then propose three algorithms for determining the best allocation of units to strata given the estimator's plannned use: Neyman allocation; a "naïve" design assuming no unmeasured confounding in the observational study; and a robust design accounting for the imperfect parameter estimates we would obtain from the observational study with unmeasured confounding. We propose guardrails on the designs, so that our experiment could be reasonably analyzed without shrinkage if desired.

We demonstrate the viability of these experimental designs through a simulation study involving a rare, binary outcome. Lastly, we deploy our methods on real data from the Women's Health Initiative, a 1991 study estimating the health effects of hormone therapy on postmenopausal women. In particular, we determine how many units should be allocated to each treatment arm in each stratum of interest in order to maximally reduce estimation risk given the planned use of the shrinkage estimator. We find improved design provides further benefits over and above the benefit of the shrinkage estimator itself.

MSC2020 subject classifications: Primary 62D20, 62L05, 62C12; secondary 65K10, 65D30.
Keywords and phrases: Causal inference, experimental design, sensitivity analysis, empirical Bayes.

Received June 2022.

1. Introduction

Recent years have seen increased interest in methods to integrate observational data with experimental data [9]. Such methods have been used to estimate average causal effects in target populations [4, 22], identify heterogeneous treatment effects [31], and improve precision in causal estimation [13].

This surge in methodological development is motivated, at least in part, by the proliferation of observational databases. Such repositories provide statisti-

cians with rich new data sources from which to learn. Yet the lurking danger of unmeasured confounding yields rightful trepidation about incorporating these data into estimation procedures [9]. In [36], the authors proposed a procedure for shrinking causal estimates for the individual strata of a stratified experiment toward the analogous estimates from an observational study. Shrinkage estimators are attractive in that they allow researchers to use observational data in tandem with experimental data, while protecting the integrity of the randomization of the experiment. Under testable conditions, they provide a guaranteed reduction in expected loss, relative to using the experimental data alone.

Separately, [38] proposed a method to design more powerful stratified experiments by utilizing information from an observational study to inform decisions about how to allocate a sample across given strata and treatment arms. Risk reductions from this method are more modest, owing to the fact that the observational data is used only for design and not for inference.

Here, we combine the approaches of design and shrinkage. In particular, if we *plan* to use shrinkage, how – given a fixed budget of units – should we allocate units across strata, and determine the proportion to treat within each stratum, in a prospective RCT to minimize estimation risk? We answer this question with an optimization: we minimize expressions for the risk of the shrinkage estimator across possible experimental designs, using information estimated from the observational study. We thus design to make our planned experiment serve as a good "complement" to the observational data, allowing for significant gains in estimation precision when eventually deploying the shrinkage estimator. However, we also want the experiment to be usable in its own right, and therefore impose guardrails on the design such that the stratum-specific conditional average treatment effect (CATE) estimates will be sufficiently precise even if we do not ultimately decide to shrink these estimates toward those obtained from the observational study.

The remainder of this paper proceeds as fellow. In Section 2, we define our problem and introduce notation and assumptions. Section 3 introduces our choice of shrinkage estimator, $\boldsymbol{\kappa}_1$, and demonstrates how to compute its risk efficiently. This section also discusses three different heuristics under which analysts can design prospective experiments with the intent of leveraging $\boldsymbol{\kappa}_1$ on the final results, while also protecting the utility of the experiment on its own. Section 4 contains two simulation studies that highlight the risk improvements that can be attained by designing toward shrinkage. In section 5 we turn to a real data application of the Women's Health Initiative (WHI), a 1991 study of the effects of hormone therapy (HT) on health outcomes for postmenopausal women. In particular, we design a small RCT to test the effect of hormone therapy (HT) on coronary heart disease using the methods of this paper, finding that our proposed design would be likely to have improved risk over designs that did not take the planned use of shrinkage estimators into account. Section 6 concludes.

2. Set-up

2.1. Notation

We operate in a stratified setting, with fixed subgroups $k = 1, \ldots, K$. The subgroups could be defined by subject matter knowledge, a consequence of a set of baseline covariates, or the result of a modern machine learning method for uncovering heterogeneous treatment effects [20, 46]. Regardless, the stratification scheme is taken to be known prior to the experimental design phase. Our goal is solely to determine sample sizes per stratum and treatment status for a future experiment on similar units. We suppose we have access to a pilot dataset obtained from an observational study.

The pilot dataset comprises n_o total units, indexed by j. With each unit, we associate two potential outcomes, $Y_j(0), Y_j(1) \in \mathbb{R}$, representing the unit's outcome in the presence and absence of treatment (see [40] for the introduction of potential outcomes and [34] for an overview). We also define a treatment variable $W_j \in \{0, 1\}$, representing whether unit j is treated or not; and a vector $\boldsymbol{X}_j \in \mathbb{R}^p$ representing measured covariates.

Stratum membership is denoted by a variable S_j where S_j is fully determined by the values of the observed covariates $\boldsymbol{X}_j$, i.e. $S_j = k \iff \boldsymbol{X}_j \in \mathcal{X}_k$ for some set of covariate values $\mathcal{X}_k$. As discussed in Assumption 2, the stratification is intended to incorporate known effect moderators. In typical cases, we expect the covariates which define stratum membership (the set of effect moderators) to be much smaller than the full set of observed covariates.

The quartets $(\boldsymbol{X}_j, W_j, Y_j(0), Y_j(1))$ are sampled i.i.d. from the observational distribution F_o. Denote as $\mathbb{E}_o$ and var_o the expectation and variance operators under F_o.

We design a future blocked experiment of the same treatment. For the experiment, we will recruit n_{rk} units for each stratum k. Within the stratum, we will then randomize n_{rkt} of those units to receive the treatment and $n_{rkc} = n_{rk} - n_{rkt}$ to receive the control. We call the set of tuples $\boldsymbol{d} = \{(n_{rkt}, n_{rkc})\}_{k=1}^{K} \in \mathbb{Z}^{K \times 2}$ the "design." We impose a total sample size constraint such that

$$\sum_k n_{rkt} + n_{rkc} = n_r,$$

for some fixed integer n_r.

Denote as F_r the sampling distribution for the triplets $(\boldsymbol{X}_i, Y_i(0), Y_i(1))$ among units i in the future experimental population. We denote the *conditional* sampling distribution for units in stratum k as $F_{r|S_i=k}$. The experiment can be understood as, for each stratum k, drawing n_{rk} units from $F_{r|S_i=k}$ and then assigning W_i by choosing a simple random sample of size n_{rkt} from the set of n_{rk} recruited units. We define as $\mathbb{E}_r$ and var_r the expectation and variance operators over both the sampling and treatment randomization in the future experiment.

2.2. Assumptions and loss function

We make the following standard assumption.

Assumption 1 (Consistency)**.** *For each unit ℓ in the pilot study or the future experiment, the observed outcome $Y_\ell \in \mathbb{R}$ is given by*

$$Y_\ell = W_\ell Y_\ell(1) + (1 - W_\ell) Y_\ell(0),$$

that is, there is only one "version" of the treatment.

A key, domain-specific assumption is that the conditional average treatment effects (CATEs) within each stratum are shared between the pilot observational and future RCT datasets, i.e.

Assumption 2 (Common Conditional Average Treatment Effects)**.** *For each stratum $k = 1, \ldots, K$,*

$$\mathbb{E}_o\left(Y(1) - Y(0) \mid S = k\right) = \mathbb{E}_r\left(Y(1) - Y(0) \mid S = k\right) \equiv \tau_k.$$

Assumption 2, sometimes called a "transportability condition" [9], imposes a congruency on the distributions F_o and F_r. This assumption gives us a common target of estimation, the vector of shared CATEs:

$$\boldsymbol{\tau} = \left(\tau_1, \ldots, \tau_K\right)^{\mathsf{T}}.$$

We will typically invoke the following slightly stronger version of Assumption 2:

Assumption 3 (Common Conditional Potential Outcome Moments)**.** *For each stratum $k = 1, \ldots, K$, and $w \in \{0, 1\}$,*

$$\begin{aligned} \mathbb{E}_o\left(Y(w) \mid S = k\right) &= \mathbb{E}_r\left(Y(w) \mid S = k\right) \quad \textit{and} \\ \mathrm{var}_o\left(Y(w) \mid S = k\right) &= \mathrm{var}_r\left(Y(w) \mid S = k\right). \end{aligned}$$

Neither Assumption 2 nor Assumption 3 can be tested prior to the design of the experiment, as data from the experimental population will not yet have been acquired at that stage. Hence, the viability of these assumptions must be informed exclusively by subject matter knowledge. In cases where the pilot and experimental populations differ markedly conditional on the stratification—e.g. if the pilot data and experimental data are collected at very different timepoints or in very different geographies—then these assumptions may be suspect. However, if the stratification meaningfully captures population heterogeneity and the pilot observational and experimental populations are comparable, these assumptions are reasonable to make.

We define

$$\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \left(\hat{\tau}_{o1}, \ldots, \hat{\tau}_{oK}\right)^{\mathsf{T}},$$

as the set of stratum causal estimates arising from the pilot study. These estimates can be obtained using a difference-in-means estimator, or a more complex

estimator. Though not yet realized, denote as $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ the vector of difference-in-means estimates from the future experiment.

Our eventual causal estimator,

$$\hat{\boldsymbol{\tau}} \equiv \boldsymbol{f}\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}, \dots\right) = \left(\hat{\tau}_1, \dots, \hat{\tau}_K\right)^{\mathsf{T}},$$

will be a function of $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, as well as other quantities estimated from the data. The target of estimation is *not* an overall average treatment effect over the experimental population. Rather, we seek to obtain good estimates simultaneously for each of the stratum CATEs, $\tau_1, \dots, \tau_K$. Under Assumption 2, we can define a loss function under which we evaluate $\hat{\boldsymbol{\tau}}$. We use the simple, unweighted L_2 loss,

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

In designing our experiment, we seek to minimize the expected L_2 loss of $\hat{\boldsymbol{\tau}}$ in estimating $\boldsymbol{\tau}$. We treat $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ as fixed, yielding the risk expression,

$$\mathcal{R}(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}}) = \mathbb{E}_r\left(\mathcal{L}(\boldsymbol{\tau}, \hat{\boldsymbol{\tau}})\right).$$

We optimize over an estimator's risk—rather than its precision—because we typically use Empirical Bayes estimators for $\hat{\boldsymbol{\tau}}$, and these estimators are not unbiased.

2.3. Related problems

We can relate the problem of optimizing our experiment to several well-studied problems in causal inference and experimental design.

Were the pilot study itself randomized, then under F_o we would have $W \perp\!\!\!\perp Y(0), Y(1) \mid X$. Then, this problem would be closely related to adaptive experimental designs. Adaptive designs refer to trials conducted in multiple phases, in which information from early phases can be used to design later phases. Such designs can encompass a wide variety of study choices [17]. In our case, we are interested in using prior information to determine how many individuals are allocated to each stratum and treatment arm. This problem has a rich history, stretching back to the work of Thompson [44, 45]. While Thompson sampling was originally designed for a more generic problem involving maximizing the expected reward, it can be used for estimation of average stratum treatment effects, as discussed in [30]. Adaptive experimental design is an area of active research, though much recent work has focused on methods that define strata in the second phase, rather than taking strata as fixed [42, 2]. For modern methods that incorporate a fixed stratification scheme (as we do in this manuscript), see [18] and [5].

Another related setting is one in which the pilot study is an observational study, but our ultimate causal estimator $\hat{\boldsymbol{\tau}}$ would simply be $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, the vector of difference-in-means estimators arising from the experiment. This problem

is closely related to the survey sampling work of Neyman [27]. Neyman computed the optimal allocation for a stratified survey—under a budget constraint—supposing pilot estimates of variance could be obtained for each stratum. These ideas can easily be extended to the causal inference setting. Under Assumption 3, pilot stratum variance estimates are obtainable from the observational study [21]. [38] demonstrates that efficiency gains are possible even if there is unmeasured confounding in the observational study, as long as we can bound the magnitude of the unmeasured confounding using a sensitivity model [43]. This is because the observational data can indicate parts of the covariate space where variation is higher or lower—and hence, where experimenters should over- or undersample.

Lastly, suppose that the pilot study were an observational study, but that the experimental study were already completed, and our goal would be to choose an estimator $\hat{\boldsymbol{\tau}}$ to trade off between causal estimates derived from the two data sources. This is an example of a "data fusion" problem [4]. Many methods rely on unconfoundedness in the observational study [37, 1]. Other papers have sought to weaken this condition, frequently utilizing alternative assumptions to proceed with merged estimations. Kallus [22] assumes that the hidden confounding has a parametric structure that can be modeled effectively. Peysakhovich and Lada [31] propose a method for when the observational data are time series and the bias preserves a unit-level rank ordering. Recent years have seen several new proposals for adaptive estimators [7, 29, 47, 6]. For an excellent overview of some of the available methods, see [9].

2.4. Principles guiding estimator choice and experimental design

In the remainder of this manuscript, we suppose the pilot is an observational study, and the experimental study has yet to be implemented. We will choose a causal estimator $\hat{\boldsymbol{\tau}} = \boldsymbol{f}\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}, \dots\right)$ for combining the observational and experimental data within each stratum. Then, we will design our experiment explicitly to minimize the risk of this estimator. Our approach is analogous to an adaptive trial, but distinct because we are using non-experimental evidence to choose the allocation at the design phase of the experiment.

Because the observational study is not randomized, we will typically conduct some form of statistical adjustment to reduce confounding bias in the observational data. In particular, we use stabilized inverse probability of treatment weighting (SIPW) as our adjustment. SIPW involves estimating the propensity score—the probability of treatment in the observational study—as a function of the observed covariates, and reweighting the units by the inverses of their estimated propensity score. For more details, see [21].

We highlight several characteristics of the shrinkage estimator and experimental design procedure that we consider ideal.

First, we would like our estimator to be *robust to unmeasured confounding* in the observational study. The assumption of unconfoundedness—roughly, that all variables affecting both the treatment probability and the outcome have

been measured in the observational study—is fundamentally untestable, and rarely holds in practice [21]. Moreover, our problem is somewhat asymmetric: a simple vector of difference-in-means estimates from the experimental study will be unbiased, and will be "good enough" in many cases. Hence, we do not want to incorporate the observational data unless we have strong guarantees that it will reduce statistical risk. Thus, our chosen estimator should not be overly susceptible to bias due to unmeasured confounding in the observational study.

One might think of using simpler data fusion estimators, based on precision-weighted convex combinations of $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, to achieve this end (see, e.g., [37]). Unfortunately, these estimators were designed under the assumption that unconfoundedness holds in the observational study. Hence, they do not exhibit the desired robustness property.

Our second criterion is that we would like our procedure to generate experiments that are still valid if they are analyzed alone. We term this feature *detachability*. To motivate this idea, consider an extreme case where our design algorithm tells us not to sample any experimental units for a given stratum, under the assumption that the observational study estimate from that stratum is sufficiently accurate. Suppose that we later learn from stakeholders that they would prefer to report causal estimates using exclusively the experimental data. In this case, we would be out of luck: the experiment cannot provide a causal estimate from the given stratum, so we must either redefine our estimand or conduct another experiment. To avoid such extreme cases, we would like to limit the space of possible designs to those that would yield reasonable estimates in the case that we choose to use $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ alone. Note also that achieving detachability will still require the approximate validity of Assumption 3, as an experiment designed under faulty pilot variance estimates may have imprecise estimates of the stratum-specific CATEs due to poor allocation of units across strata and treatment arms.

3. Designing towards shrinkage

3.1. Shrinkage estimators for the CATE

In [15] and [16], Green and Strawderman consider how to shrink between an unbiased estimator and a biased estimator. Their goal is to derive Empirical Bayes estimators that guarantee a risk reduction relative to using the unbiased estimator alone. The problem turns out to be quite similar to James-Stein estimation.

In this paper, we primarily consider $\boldsymbol{\kappa}_1$, a shrinkage estimator introduced in [36], which builds on the work of Green and Strawderman. In particular, let $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \in \mathbb{R}^K$ be the unbiased (RCT) estimator of the K strata CATEs, and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} \in \mathbb{R}^K$ be the corresponding biased (observational study) estimator. Denote as $\boldsymbol{\Sigma}_r = \text{diag}(\sigma^2_{rk}) \in \mathbb{R}^{K \times K}$ the diagonal covariance matrix of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$; the square root of the diagonal of $\boldsymbol{\Sigma}_r$ would be the standard errors for the strata-level

estimates $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$. Under mild conditions, we can assume $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} \sim \mathcal{N}(\boldsymbol{\tau}, \boldsymbol{\Sigma}_r)$ [see e.g. 26], where $\boldsymbol{\tau}$ is the vector of true causal estimates. We primarily focus on

$$\boldsymbol{\kappa}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \left(\frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right) (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) .$$

$\boldsymbol{\kappa}_1$ shrinks each component of the unbiased estimator toward its counterpart in the biased estimator by the same multiplicative factor in parentheses. This estimator is intuitive to understand, and it often outperformed competitor estimators—including those proposed by Green and Strawderman—in simulations based on data from the Women's Health Initiative [36]. We discuss the use of alternative estimators, such as those that differentially shrink each component of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ toward its corresponding entry in $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$, in the Supplementary Material.

Using results from [41], the risk of $\boldsymbol{\kappa}_1$ is:

$$\mathcal{R}(\boldsymbol{\kappa}_1) = \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{K} \left(1 + \mathbb{E}_r \left(\frac{4(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^{\mathsf{T}} \boldsymbol{\Sigma}_r (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}{((\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}))^2} - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})} \right) \right), \tag{1}$$

where risk is defined as $\mathbb{E}_r(\mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\kappa}_1))$, and $\mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\kappa}_1) = \|\boldsymbol{\kappa}_1 - \boldsymbol{\tau}\|^2$. The expectation is with respect to $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ only. $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ is treated as a constant vector: we are interested in the risk of our future experiment, given the data we have up to the point of planning that experiment.

As shown in [36], the estimator is guaranteed to dominate $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under the squared-error loss as long as the condition

$$4 \max_k \sigma_{rk}^2 < \sum_k \sigma_{rk}^2 \tag{2}$$

is satisfied. In other words, if Condition 2 holds, $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ is inadmissible with respect to squared error risk.

The $\boldsymbol{\kappa}_1$ estimator's shrinkage is an estimated expression that could be negative, which would push $\boldsymbol{\kappa}_1$ *away* from the randomized trial estimates. We can instead truncate negative shifts at 0. This positive part analogue of $\boldsymbol{\kappa}_1$ is

$$\boldsymbol{\kappa}_{1+} = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(1 - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}} (\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})} \right)_+ (\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) ,$$

which constrains the shrinkage estimator from "over-adjusting." The risk of $\boldsymbol{\kappa}_{1+}$ is guaranteed to be strictly lower than that of $\boldsymbol{\kappa}_1$. In the simulations of [36], $\boldsymbol{\kappa}_{1+}$ routinely dominated $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ even when Condition 2 was not met.

3.2. Exact risk calculation under known parameters

Our goal will be to optimize the experimental design over the risk given in Expression (1). Define $\boldsymbol{\xi}$ as the (negative) error of the observational study:

$$\boldsymbol{\xi} = \tau - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} = \mathbb{E}_r(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}) - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}.$$

This error is both the bias of the observational study and any stochastic error; we are conditioning on both these things. Nonetheless, for simplicity of exposition, we will often refer to $\boldsymbol{\xi}$ as the "bias vector."

We can then express the risk in Expression (1) as the expectation of a ratio of quadratic forms of a multivariate normal variable centered at $\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi}$, assuming $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\xi}$ are both known:

$$\mathcal{R}(\boldsymbol{\kappa}_1) = \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{K}\left(1 + \mathbb{E}\left(\frac{4\cdot \boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^2\boldsymbol{\nu}}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r\boldsymbol{\nu})^2} - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r)}{\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r\boldsymbol{\nu}}\right)\right),$$

where $\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi}, \boldsymbol{I}_K)$.

Exact integral expressions for the above components can be found in [3]. In particular,

$$\begin{aligned}
\mathbb{E}\left(\frac{\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^2\boldsymbol{\nu}}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r\boldsymbol{\nu})^2}\right) &= \int_0^\infty \det(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1/2}\cdot \\
&\qquad \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1/2}(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\xi}\right)\right)\cdot \\
&\qquad \left(\operatorname{tr}(\boldsymbol{R}) + (\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})^\mathsf{T}\boldsymbol{R}(\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})\right) t dt, \text{ and} \\
\mathbb{E}\left(\frac{1}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r\boldsymbol{\nu})}\right) &= \int_0^\infty \det(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1/2}\cdot \\
&\qquad \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1/2}(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\xi}\right)\right) dt
\end{aligned} \tag{3}$$

where

$$\begin{aligned}
\boldsymbol{L} &= (\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1/2}, \qquad \text{and} \\
\boldsymbol{R} &= \boldsymbol{L}^\mathsf{T}\boldsymbol{\Sigma}_r^2\boldsymbol{L}.
\end{aligned}$$

The integrals in Expression (3) can be computed via numerical integration, yielding an efficient evaluation of the risk for each possible choice of the parameter values.

3.3. Variance estimation

The exact risk calculation discussed in the prior subsection relies on knowledge of $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}_r$. We use the observational study to motivate plausible values for these quantities, and then optimize the risk over allowed sets of such values.

We determine $\boldsymbol{\Sigma}_r$ by expressing it in terms of strata-specific variances, following classic analysis of randomized trials. In particular, for a standard difference-in-means estimator for each stratum-specific CATE, the entries in $\boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_{rk}^2)$ are functions of the RCT design $\boldsymbol{d} = \{(n_{rkt}, n_{rkc})\}_{k=1}^K$ as well as the stratum-specific potential outcome variances $\boldsymbol{V} = \left\{\left(\sigma_{rkt}^2, \sigma_{rkc}^2\right)\right\}_{k=1}^K$ via the standard

relation

$$\sigma^2_{rk} = \frac{\sigma^2_{rkt}}{n_{rkt}} + \frac{\sigma^2_{rkc}}{n_{rkc}}.$$

Hence, if we can obtain a reasonable estimate of $\boldsymbol{V}$, we have a reasonable estimate of $\boldsymbol{\Sigma}_r$.

To estimate $\boldsymbol{\Sigma}_r$, we make an appeal to our assumptions. In the abstract, Assumption 3 is strong. However, in the settings in which our data-combination methods are useful – that is, settings where we assume strong congruency between the observational and experimental data, such that Assumption 2 holds approximately – Assumption 3 is often plausible. Hence, to estimate $\boldsymbol{V}$, we suppose Assumption 3 and unconfoundedness both hold, and then use the observational study to estimate the variance of the potential outcomes in each stratum. It is possible that residual unmeasured confounding may induce some bias in the estimation, even after statistical adjustment. However, while the magnitude of such biases is a concern for precise causal point estimation, it is typically small enough as to not pose a major challenge for pilot variance estimation.

Denote the strata-specific variance estimates obtained from the observational data as

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

We plug these into our expression for $\boldsymbol{\Sigma}_r$ to obtain an estimate of this quantity for any value of the design $\boldsymbol{d}$.

The bias vector $\boldsymbol{\xi}$ cannot be estimated before experimental data is collected, as we would not have any form of "ground truth" against which to compare $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. That being said, if our observational study is large, our statistical adjustment strategy is sound, and our selection on observables assumptions hold, we would expect $\boldsymbol{\xi}$ to be small. In the next sections we offer heuristics for how to proceed under uncertainty regarding $\boldsymbol{\xi}$ and the estimated $\boldsymbol{\Sigma}_r$.

3.4. Design options

In this section, we consider several heuristics for designing the experiment in the absence of perfect knowledge of $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\xi}$.

3.4.1. Neyman allocation

A simple approach to the experimental design problem is to assume that there is no residual unmeasured confounding in the observational study. Importantly, we make the unconfoundedness assumption for the purposes of *design* only. The unconfoundedness assumption need not be strictly true to ensure good performance of $\boldsymbol{\kappa}_1$ once our experiment is completed. The shrinkage properties of $\boldsymbol{\kappa}_1$ ensure that its risk will be lower than $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ as long as Condition 2 is met, irrespective of the presence of residual confounding in the observational study

or error in estimating $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. Hence, we retain the implicit guarantee against a risk increase when it comes to *estimation*, even if this assumption turns out to be incorrect.

The simplest design heuristic is to then use a Neyman allocation [40] without a cost constraint, plugging in our observational-study based estimates of the variances, e.g.

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

Such a design would be optimal if the risk of our estimator were only $\text{tr}(\boldsymbol{\Sigma}_r)/K$, the first term in Expression (1). Though the design does not directly optimize over the shrinkage portion of the risk, it serves as a reasonable starting point for the purposes of design. As we will see in Section 4, it also typically yields good performance for $\boldsymbol{\kappa}_1$ in simulations.

3.4.2. Heuristic optimization assuming $\boldsymbol{\xi} = \mathbf{0}$

Per the discussion in Section 3.2, we can compute the risk exactly if both $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\xi} = \mathbb{E}_r(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}) - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ are known. Under Assumption 3 and unconfoundedness, $\boldsymbol{\Sigma}_r$ can be estimated unbiasedly for any choice of $\boldsymbol{d} = \{(n_{rkt}, n_{rkc})\}_{k=1}^K$. However, $\boldsymbol{\xi}$ may be nonzero even if unconfoundedness holds, because we consider $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ to be a fixed draw from the observational distribution, rather than a random variable.

In this section, we make the additional assumption that $\boldsymbol{\xi} = 0$. This is again an assumption of convenience: if we are wrong, our design is possibly suboptimal, but our future analysis would still be valid and we could still achieve benefit over other default designs such as equal allocation. Given $\boldsymbol{\xi}$, we have all the necessary parameter estimates to optimize $\mathcal{R}(\boldsymbol{\kappa}_1)$ over the choice of $\boldsymbol{d}$ This problem is encoded in Optimization Problem 4:

$$\begin{aligned} \text{minimize} \quad & \mathcal{R}(\boldsymbol{\kappa}_1) \\ \text{subject to} \quad & \sigma_{rk}^2 = \frac{\hat{\sigma}_{kt}^2}{n_{rkt}} + \frac{\hat{\sigma}_{kc}^2}{n_{rkc}}, \quad k = 1, \ldots, K, \\ & 0 < n_{rkt}, n_{rkc}, \quad k = 1, \ldots, K, \\ & n_r = \sum_k n_{rkt} + n_{rkc}, \end{aligned} \tag{4}$$

Unfortunately, $\mathcal{R}(\boldsymbol{\kappa}_1)$ is not convex in d. However, Optimization Problem 4 can be approximately solved using a greedy algorithm. Define

$$\boldsymbol{d}_j = \{(n_{rkt}^{(j)}, n_{rkc}^{(j)})\}_k \in \mathbb{Z}^{K \times 2}$$

as the allocation of RCT units to strata and treatment level at iteration j of the algorithm. Next, define

$$D_j = \{\boldsymbol{d} \in \mathbb{Z}^{K \times 2} \mid \boldsymbol{d} \text{ swaps exactly 1 unit across strata or treatment from } \boldsymbol{d}_j\}.$$

Because there are K strata and two treatment levels, the "swap set" D_j will contain

$$2\binom{2K}{2} = 2K \times (2K-1)$$

possible allocations. Some of these allocations may correspond to invalid designs, as will be discussed in Section 3.5.

Define $\mathcal{R}_1(\boldsymbol{d}, \boldsymbol{V}, \boldsymbol{\xi})$ as the value of $\mathcal{R}(\boldsymbol{\kappa}_1)$ evaluated under the design $\boldsymbol{d}$ with estimated stratum potential outcome variances $\boldsymbol{V}$ and bias vector $\boldsymbol{\xi}$. We will evaluate $\mathcal{R}_1$ under estimated variances $\hat{\boldsymbol{V}} = \left\{\left(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc}\right)\right\}_{k=1}^{K}$ and $\boldsymbol{\xi} = \mathbf{0}$.

Now, we can approximately solve Optimization Problem 4 by running Algorithm 5:

$$\begin{aligned}
&\texttt{Start with design } \boldsymbol{d}_0 = \{(n^{(0)}_{rkt}, n^{(0)}_{rkc}\}_k\,.\\
&\texttt{For iteration } j = 1, 2, \ldots:\\
&\quad \texttt{For each design } \boldsymbol{d} \texttt{ in } D_{j-1}\texttt{:}\\
&\qquad \texttt{Compute } \mathcal{R}_1(\boldsymbol{d}, \hat{\boldsymbol{V}}, \mathbf{0})\,.\\
&\quad \texttt{Set } \boldsymbol{d}_j = \underset{\boldsymbol{d} \in D_{j-1}}{\arg\min}\, \mathcal{R}_1(\boldsymbol{d}, \hat{\boldsymbol{V}}, \mathbf{0})\\
&\quad \texttt{If } \mathcal{R}_1(\boldsymbol{d}_j, \hat{\boldsymbol{V}}, \mathbf{0}) >= \mathcal{R}_1(\boldsymbol{d}_{j-1}, \hat{\boldsymbol{V}}, \mathbf{0})\\
&\qquad \texttt{Return } \boldsymbol{d}_{j-1}\,.
\end{aligned} \tag{5}$$

In words, Algorithm 5 will continue to swap units between strata and treatment levels until no swap will further reduce the estimated risk of the shrinkage estimator. The algorithm naturally enforces the sample size constraint and ensures that the returned values will be integers.

If the number of strata K is large, then Algorithm 5 may be slow to search the space of possible allocations, and hence slow to converge. An alternative approach can be found in projected gradient descent. This can be implemented via Algorithm 6,

$$\begin{aligned}
&\texttt{Start with design } \boldsymbol{d}_0 = \{(n^{(0)}_{rkt}, n^{(0)}_{rkc}\}_k\,.\\
&\texttt{For iteration } j = 1, 2, \ldots:\\
&\quad \texttt{Compute } \boldsymbol{g} = \nabla_d \mathcal{R}_1(\boldsymbol{d}_{j-1}, \hat{\boldsymbol{V}}, \mathbf{0})\\
&\quad \texttt{Set } \boldsymbol{d}_j = \texttt{Proj}\left(\boldsymbol{d}_{j-1} - \gamma \cdot \boldsymbol{g}\right)\\
&\quad \texttt{If } \mathcal{R}_1(\boldsymbol{d}_j, \hat{\boldsymbol{V}}, \mathbf{0}) >= \mathcal{R}_1(\boldsymbol{d}_{j-1}, \hat{\boldsymbol{V}}, \mathbf{0})\\
&\qquad \texttt{Return } \boldsymbol{d}_{j-1}\texttt{,}
\end{aligned} \tag{6}$$

where γ is a learning rate and $\texttt{Proj}(\cdot)$ represents a projection onto the constraint set. Practically speaking, $\nabla_d \mathcal{R}_1(\boldsymbol{d}_{j-1}, \hat{\boldsymbol{V}}, \mathbf{0})$ (the gradient of the shrinker risk with respect to the design) can be computed via numerical differentiation. Projection onto the constraint set can be closely approximated by a two-step process. First, we find the point in $\mathbb{R}^{K\times 2}$ nearest to $\boldsymbol{d}_{j-1} - \gamma \cdot \boldsymbol{g}$ which satisfies

the non-negativity and sample size constraints; this is a simple convex optimization problem with affine constraints. Second, we take this projected point and round up its components in descending order of the decimal term, such that the resulting sample sizes are integers and the sum of the rounded sample sizes is preserved at exactly n_r.

These two algorithms can also be composed: we run Algorithm 6 until a stopping condition is reached, and then run Algorithm 5 until convergence. This approach yields significant speed improvements, since the gradient descent method of Algorithm 6 rapidly gets close to an optimum, and the greedy approach of Algorithm 5 quickly converges to it. This compound algorithm is used in the simulation study in Section 4.

Because the objective is non-convex, it is plausible that Algorithm 5, Algorithm 6, or their composition could get stuck at local optima. Practically, we recommend running the algorithm at a few different starting points (e.g. an equally allocated design, the Neyman allocation, and several randomly chosen designs), and choosing the design which achieves the minimum value of the risk. While this approach is not guaranteed to find the global optimum, it will nonetheless find a point with a reduced value of the objective function. Each step will consistently find an improved design. In particular, if the relevant assumptions hold, the best found local minimum will be guaranteed to improve efficiency relative to a Neyman allocation.

3.4.3. Heuristic optimization assuming worst-case error under $\boldsymbol{\Gamma}$*-level unmeasured confounding*

The assumption that $\boldsymbol{\xi} = \mathbf{0}$ is fundamentally optimistic: it is unlikely to hold even in the absence of unmeasured confounding. If there are unmeasured confounders, it may be far from the truth. We can take a more defensive approach by imposing a sensitivity model on the observational study, and optimizing under the worst-case choice of $\boldsymbol{\xi}$.

In particular, we constrain the magnitude of the unmeasured confounding by imposing the marginal sensitivity model of Tan [43]. Under this model, a key odds ratio—between the treatment probability conditional on the potential outcomes and covariates and the treatment probability conditional on covariates only—is bounded between $1/\Gamma$ and Γ, for a user-chosen parameter $\Gamma \geq 1$. The Tan model can be seen as extending the popular Rosenbaum sensitivity model [35] to the setting of inverse probability weighting. Practical methods for calibrating the choice of Γ, using observed covariates, can be found in [11] and [23], among others.

Under Assumption 3, we have

$$\begin{aligned} \mu_{kt} &\equiv \mathbb{E}_o(Y(1) \mid S = k) = \mathbb{E}_R(Y(1) \mid S = k), \qquad \text{and} \\ \mu_{kc} &\equiv \mathbb{E}_o(Y(0) \mid S = k) = \mathbb{E}_R(Y(0) \mid S = k), \end{aligned}$$

for $k = 1, \ldots, K$. For any choice of Type I error bound $\alpha \in (0, 1)$, we can use the method of Zhao, Small, and Bhattacharya [48] to obtain intervals $\left(\ell_{kt}^{(\Gamma,\alpha)}, u_{kt}^{(\Gamma,\alpha)}\right)$

and $\left(\ell_{kc}^{(\Gamma,\alpha)}, u_{kc}^{(\Gamma,\alpha)}\right)$ such that parameters μ_{kt} and μ_{kc} reside within the intervals with at least $1-\alpha$ probability as long as the true confounding structure lies within the sensitivity model parameterized by Γ. The method relies on convex optimization and the bootstrap in order to generate valid confidence sets.

Per the results in [38], if the outcome of interest is binary, we can use the confidence sets on μ_{kt} and μ_{kc} to obtain valid confidence sets for the stratum potential outcome variances σ_{kt}^2 and σ_{kc}^2 via the relations

$$\sigma_{kt}^2 = \mu_{kt} \cdot (1-\mu_{kt}) \qquad \text{and} \qquad \sigma_{kc}^2 = \mu_{kc} \cdot (1-\mu_{kc}).$$

For a full justification, see [38].

Putting these ideas together, we obtain a robust approach that can be utilized for any problem with binary outcomes. Under our calibrated choice of Γ and a reasonable choice of α, we set for each $k = 1, \ldots, K$,

$$\xi_k = \max\left(\left|u_{kt}^{(\Gamma,\alpha)} - \ell_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right|, \left|\ell_{kt}^{(\Gamma,\alpha)} - u_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right|\right),$$

the worst-case value of the error under our sensitivity model. We collect these quantities into a vector $\tilde{\boldsymbol{\xi}}_{\boldsymbol{\Gamma}}$. Next, we collect the corresponding values of the variances, e.g.

$$\tilde{\sigma}_{kt}^2 = \begin{cases} u_{kt}^{(\Gamma,\alpha)} \cdot (1-u_{kt}^{(\Gamma,\alpha)}) & \text{if } \left|u_{kt}^{(\Gamma,\alpha)} - \ell_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right| > \left|\ell_{kt}^{(\Gamma,\alpha)} - u_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right| \\ \ell_{kt}^{(\Gamma,\alpha)} \cdot (1-\ell_{kt}^{(\Gamma,\alpha)}) & \text{otherwise} \end{cases}$$

and

$$\tilde{\sigma}_{kc}^2 = \begin{cases} \ell_{kc}^{(\Gamma,\alpha)} \cdot (1-\ell_{kc}^{(\Gamma,\alpha)}) & \text{if } \left|u_{kt}^{(\Gamma,\alpha)} - \ell_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right| > \left|\ell_{kt}^{(\Gamma,\alpha)} - u_{kc}^{(\Gamma,\alpha)} - \hat{\tau}_{ok}\right| \\ u_{kc}^{(\Gamma,\alpha)} \cdot (1-u_{kc}^{(\Gamma,\alpha)}) & \text{otherwise} \end{cases}$$

into a matrix $\tilde{\boldsymbol{V}}_{\boldsymbol{\Gamma}}$.

Finally, we can evaluate our function $\mathcal{R}_1(\boldsymbol{d}, \tilde{\boldsymbol{V}}_{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\xi}}_{\boldsymbol{\Gamma}})$ to obtain the risk of $\boldsymbol{\kappa}_1$ for any experimental design $\boldsymbol{d}$ under these parameters. The procedure is henceforth analogous to the one used in the prior section: we run Algorithm 5 or Algorithm 6, substituting $\mathcal{R}_1(\boldsymbol{d}, \tilde{\boldsymbol{V}}_{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{\xi}}_{\boldsymbol{\Gamma}})$ for $\mathcal{R}_1(\boldsymbol{d}, \hat{\boldsymbol{V}}, \boldsymbol{0})$, and obtain the design that yields the lowest value of the risk.

The approach does not readily generalize to continuous outcomes. This is because, if $Y_i(0), Y_i(1) \in \mathbb{R}$, then the potential outcome variances corresponding to the worst-case value of the error under the sensitivity model are not a simple function of the potential outcome bounds. Hence, it is not immediately clear how to populate the matrix $\tilde{\boldsymbol{V}}_{\boldsymbol{\Gamma}}$ under sensitivity parameter Γ. We highlight this challenge as an opportunity for future work.

3.5. Imposing guardrails on designs

We discuss three plausible constraints that can be incorporated on the set of possible designs. Generally, we suggest imposing a *minimum sample size constraint* such that the allocations to any stratum and treatment group cannot

be lower than some value $\text{SS}_{\min}$. This constraint serves two purposes. First, the risk expression in Expression (1) depends explicitly on the normality of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$. In simulations, we find that this expression is still quite accurate under modest deviations from normality. Nonetheless, if sufficiently few units are allocated to any stratum or treatment arm, a Central Limit Theorem need not hold even approximately. A minimum sample size constraint averts this problem. Second, this constraint naturally helps improve *detachability*, because it prevents the variance of any entry of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ from growing too large.

Second, we suggest imposing an explicit detachability constraint on top of the sample size constraint. The goal is to ensure that an analyst could analyze the experiment on its own if case stakeholders ultimately decide not to use the observational data. One constraint that can achieve this purpose is to first select some baseline design $\tilde{\boldsymbol{d}} = \{\tilde{n}_{rkt}, \tilde{n}_{rkc}\}_k$, e.g. equal allocation or Neyman allocation, and then consider only designs $\boldsymbol{d} = \{n_{rkt}, n_{rkc}\}_k$ for which

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

where $\delta_d \geq 1$ is a user-chosen tolerance parameter. In words, this approach restricts the set of possible designs to those designs $\boldsymbol{d}$ for which the estimated risk of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ exceeds the estimated risk of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under a default design $\tilde{\boldsymbol{d}}$ by a multiplicative factor no larger than δ_d (all assuming the observational study point estimates of the strata potential outcome variances are correct).

In some cases, one may also want to impose a *risk reduction* constraint. Imposing this constraint means that we consider only those designs that are estimated to satisfy Condition 2, i.e. designs $\boldsymbol{d} = \{n_{rkt}, n_{rkc}\}_k$ for which

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

Unlike the minimum sample size and detachability constraints, the risk reduction constraint may not be appropriate in many cases. Its imposition guarantees that the estimator $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ will be inadmissible relative to $\boldsymbol{\kappa}_{1+}$, which may give greater confidence to analysts in using the shrinkage estimator. However, the condition cannot be satisfied if there are fewer than five strata, and may be unnecessarily restrictive if the potential outcome variances are highly heteroscedastic across strata. Moreover, we find in simulation that $\boldsymbol{\kappa}_{1+}$ very frequently achieves risk reductions relative to $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ even in cases when Condition 2 is not met. Thus, this constraint should be imposed with caution.

Algorithm 5 can easily incorporate each of these three constraints. When evaluating the risk of any potential design $\boldsymbol{d}$ in the swap set D_{j-1}, one can check if it satisfies the constraint and, if not, set the risk equal to infinity. This will naturally force the algorithm to choose among designs that satisfy the constraints. Algorithm 6 can incorporate the minimum sample size and detachability constraints in the projection step, as the former is an affine inequality and the latter a convex inequality. However, the risk reduction constraint cannot be easily incorporated into Algorithm 6, as it involves an inequality between two convex

functions of the design $\boldsymbol{d}$ and hence is not compliant with disciplined convex programming [14].

In the Supplementary Material, we discuss alternative versions of the detachability and risk reduction constraints, which seek to incorporate greater robustness to unmeasured confounding.

4. Simulation study

4.1. Simulation set-up

We pattern our simulations on those in [37], considering a situation with a relatively rare, binary outcome and a modest effect size.

We first generate an observational super-population of 1×10^6 units, and an experimental super-population of the same size. We suppose that the completed observational study comprises 20,000 units, sampled a single time from the corresponding super-population. The prospective RCT comprises 1,000 units, which will be sampled repeatedly from the experimental super-population.

We define $j \in \mathcal{O}$ as the indexing variable for the observational super-population and $i \in \mathcal{E}$ as the indexing variable for the experimental super-population. We use ℓ as an index over both populations. For each unit $\ell \in \mathcal{O} \cup \mathcal{E}$, we suppose there is a covariate vector $\boldsymbol{X}_\ell \in \mathbb{R}^5$ where $\boldsymbol{X}_\ell \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma}$ such that each covariate has unit variance and roughly a quarter of the covariances are $+0.1$, roughly a quarter are -0.1, and the remainder are 0. Such a covariance structure is roughly consistent with the applied data analysis from the Women's Health Initiative, as used in [37].

The untreated potential outcomes $Y_\ell(0)$ are sampled as independent Bernoulli random variables with

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

where α is chosen such that the average incidence rate is 10%. The treatment variables in the observational study are independent Bernoulli random variables with

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

Because $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ point in similar directions, we see strong selection bias in the observational study. In particular, treated units are much likelier to have large untreated potential outcomes.

For both datasets, we suppose the data is split into twelve strata based on the first and second covariates, which are assumed to define meaningful subgroups for which there is substantive interest in obtaining the CATEs. The strata boundaries are defined by the 25th and 50th quantiles of the first covariate and by the quartiles of the second covariate. We seek to obtain estimates of the average causal effect in each of the resultant strata.

In both datasets, we assign individual treatment effects to match three different structures. We number the strata from $k = 1, \ldots, 12$ such that stratum 1 corresponds to the lowest stratum on both covariates, stratum 2 corresponds to the lowest stratum on the first covariate and the second quartile of the second covariate, stratum 3 corresponds to the lowest stratum on the first covariate and the third quartile of the second covariate, etc. We consider three different treatment effect models. The values of τ_k, the stratum average treatment effects, in the constant, linear, and quadratic strata-level treatment effect models are

$$\tau_k = T, \quad \tau_k = -T \times \frac{k}{K}, \quad \text{and} \quad \tau_k = T \times \left(\frac{k}{K}\right)^2 \tag{7}$$

respectively. In each case we choose the scale $T > 0$ so that Cohen's d [8] in the observational study precisely equals 0.5 marginally across the entire population. Cohen calls this a "medium" effect size.

We generate $Y(1)$ as a binary potential outcome as follows: set the initial values of $Y_\ell(1)$ to $Y_\ell(0)$ for all units. Then randomly select $\tau_k \times n_k$ units for which $Y_\ell(0) = 0$, and set $Y_\ell(1) = 1$ for those units, where n_k is the number of units in stratum k in the super-population. Because we use the same process across both super-populations, Assumption 3 holds: the the observational and RCT data distributions have the same stratum-specific causal effects and potential outcome means and variances.

The observational data is sampled a single time. Next, leveraging the observational sample, we compute the allocations of units to strata in the RCT (the RCT "design") using each of the methods discussed in the prior section. Our optimization approach is the composition approach discussed in Section 3.4.2. Namely, we run gradient descent (Algorithm 6) until we see no improvement, and then run the greedy swapping approach (Algorithm 5) until convergence. For each of our optimized designs, we provide seven starting points to the algorithm: the equal allocation, the Neyman allocation, and five randomly chosen starting points.[1] Happily, we find that the starting point makes no difference: we always convergence to precisely the same design, regardless of starting point, in every tested condition in the simulations. This increases confidence that the optimization approach is not very sensitive to the starting point for the algorithm.

Once we have each of the designs computed, we can simulate actual experiments. Under each design, we sample the RCT units from the super-population 25,000 times. For each iteration, for each stratum k, we assume treatment is assigned via a simple random sample of n_{rkt} units out of the $n_{rkt} + n_{rkc}$ units recruited for the stratum. Once the units are drawn and treatments are assigned, we compute the estimators $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, $\boldsymbol{\kappa}_1$, and $\boldsymbol{\kappa}_{1+}$. We compute the L_2 distance between each estimate and the true treatment effects τ, and take the average over all 25,000 simulations.

[1] The random starting points are generated by sampling $2 \times K$ values from a Uniform$(0, 1)$ distribution; normalizing by their sum to turn them into proportions; scaling the proportions by n_r; and then projecting the resultant values onto our constraint set such that the final values are integers and respect the minimum sample size constraint.

TABLE 1

Risk over 25,000 *iterations of* $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_1$, *and* $\boldsymbol{\kappa}_{1+}$ *under various experimental designs, in the case of no unmeasured confounding in the observational study. Risks are expressed as a percentage of the risk of* $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ *using an equally allocated experiment, for each of the three treatment effect models. The minimum non-oracle risk in each row is denoted with an underline.*

					Max Bias, Γ Value				
Est.	**Trt.**	**Equal**	**Neyman**	**Naïve**	**1.0**	**1.1**	**1.2**	**1.5**	**Oracle**
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$		100%	<u>85%</u>	89%	89%	89%	90%	91%	87%
$\boldsymbol{\kappa}_1$	c	45%	38%	<u>36%</u>	37%	37%	37%	37%	36%
$\boldsymbol{\kappa}_{1+}$		32%	28%	<u>28%</u>	28%	28%	28%	28%	28%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$		100%	<u>91%</u>	93%	94%	95%	94%	97%	92%
$\boldsymbol{\kappa}_1$	ℓ	50%	<u>41%</u>	42%	43%	42%	43%	44%	42%
$\boldsymbol{\kappa}_{1+}$		37%	34%	<u>34%</u>	35%	35%	35%	35%	34%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$		100%	87%	<u>85%</u>	90%	91%	92%	94%	85%
$\boldsymbol{\kappa}_1$	q	42%	34%	<u>33%</u>	34%	35%	36%	36%	31%
$\boldsymbol{\kappa}_{1+}$		26%	24%	<u>23%</u>	24%	24%	24%	24%	23%

The code to implement these simulations, along with a README file, is provided as Supplementary Material to this manuscript.

4.2. Ideal case: no unmeasured confounding

We begin with the simplest case: we suppose all of the covariates are measured in the observational study, so there is no residual unmeasured confounding. This is an idealized case, in which all of the selection bias in the observational study can be removed with a statistical adjustment. We fit a propensity score to the observational study data and use stabilized inverse probability of treatment weighting (SIPW) to compute the observational causal estimates.

In Table 1, we show the average L_2 errors over 25,000 simulations. We consider the equal allocation and Neyman allocation designs, as well as the "naïve" design assuming $\boldsymbol{\xi} = 0$. We also consider the "worst case" defensive approach discussed in Section 3.4.3, and compute the optimal design under errors computed with the Tan sensitivity model parameter set to 1.0, 1.1, 1.2, and 1.5. Lastly, we consider an "oracle" design, in which we run our composed optimization algorithm, but provide it with the true values of the potential outcome variances and the error in the observational study. Results are given for the three different treatment effect models: constant (c), linear (ℓ), and quadratic (q). Risk estimates are expressed as percentages of the risk of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ when using an equal allocation for the given treatment effect model.

Across the three treatment effect models, we see that results are relatively consistent in the ordering of the estimators. When using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ alone, the Neyman and naïve allocations typically perform similarly, realizing a roughly 10-15% error reduction relative to the equal allocation. The other allocations achieve slightly more modest error reductions when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$.

If we stick with the Neyman allocation but switch to using either the shrinkage estimator $\boldsymbol{\kappa}_1$ or its positive part analogue $\boldsymbol{\kappa}_{1+}$, we can realize massive additional error reductions on the order of 60 to 75%, depending on the treatment

effect model. However, the naïve allocation is typically slightly better. It performs the best among all realizable (i.e. non-oracle) designs when using $\boldsymbol{\kappa}_{1+}$ under all three treatment effect models, and when using $\boldsymbol{\kappa}_1$ in the constant and quadratic treatment effect models.

The robust allocations typically perform nearly as well as the Neyman and naïve allocations, achieving risk reductions a point or two higher. These allocations may pay a small penalty for robustness—and hence, cannot outperform the naïve approach when there is no unmeasured confounding.

4.3. Practical case: unmeasured confounding

The assumption of unconfoundedness is not defensible in many practical settings, in which a relatively sparse set of covariates are measured in the observational study. Moreover, the assumption is not testable. Hence, we run a second set of simulations in which we induce confounding by assuming that the third entry in $\boldsymbol{X}_j$ is not measured. This third covariate affects both treatment probabilities and outcomes, so bias in the observational study can no longer be fully corrected with a propensity score adjustment.

Results are given in Table 2. Notably, the residual bias in the observational study attenuates the risk reduction achievable through shrinkage and design. Relative to the risk of using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under an equally allocated experiment, our best shrinker and design combinations can realize risk reductions of about 43% in the constant treatment effect model (vs. about 72% when there was no unmeasured confounding); about 27% in the linear treatment effect model (vs. 66% with no unmeasured confounding); and about 57% under the quadratic treatment effect model (vs. 77% with no unmeasured confounding).

Optimal performers differ slightly from the simulations assuming no unmeasured confounding. Under the constant treatment effect model, we find that the robust design under $\Gamma = 1.0$ perform the best when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_1$, and $\boldsymbol{\kappa}_{1+}$. Under the linear treatment effect model, the best performer is the naïve design when using all three estimators. And under the quadratic treatment effect model, the best performer is the Neyman design when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ and the naïve design when using the shrinkers. More generally, under all treatment conditions, the Neyman, naïve, and robust designs under $\Gamma = 1$ all typically perform well.

These results point to a few practical guidelines for designing toward shrinkage. First, we find that the naïve allocation is quite robust, even when $\boldsymbol{\xi}$ is far from zero. This is evident from the fact that the naïve allocation always yields significant performance gains over the equal allocations when using a shrinkage estimator, even when unmeasured confounding is present.

Second, in the presence of unmeasured confounding, one may sometimes achieve modest further improvements from enforcing a robust design at a relatively low value of Γ. In this example, we have *not* constructed the confounding such that it matches the form of our sensitivity model. Exclusion of the third covariate induces enormous discrepancies in the treatment odds between the true and estimated propensity scores in a small proportion of individuals: for

TABLE 2

Risk over 25,000 *iterations of* $\hat{\tau}_r, \kappa_1$, *and* κ_{1+} *under various experimental designs, in the case of unmeasured confounding in the observational study via failure to measure the third covariate. 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. The minimum non-oracle risk in each row is denoted with an underline.*

					Max Bias, Γ Value				
Est.	**Trt.**	**Equal**	**Neyman**	**Naïve**	**1.0**	**1.1**	**1.2**	**1.5**	**Oracle**
$\hat{\tau}_r$		100%	91%	92%	91%	94%	94%	98%	90%
κ_1	c	69%	60%	60%	59%	61%	61%	63%	58%
κ_{1+}		64%	58%	58%	57%	59%	59%	61%	57%
$\hat{\tau}_r$		100%	93%	93%	93%	95%	97%	99%	89%
κ_1	ℓ	82%	74%	73%	74%	75%	75%	78%	70%
κ_{1+}		81%	74%	73%	74%	75%	75%	77%	70%
$\hat{\tau}_r$		100%	88%	90%	91%	92%	92%	97%	87%
κ_1	q	58%	48%	48%	49%	50%	50%	51%	47%
κ_{1+}		47%	43%	43%	43%	44%	44%	45%	42%

about 0.2% of the super-population units, the difference exceeds a multiplicative factor of 100. For most of the population, however, the true and estimated treatment odds differ by a much smaller factor. The Tan model imposes a worst-case bound on the deviation between the true and estimated odds of treatment, so no value of Γ between 1.0 and 2.0 is large enough to account for our most extreme deviations.

Nonetheless, the magnitude of the worst-case error under the Tan model correlates reasonably well with the true values of $\boldsymbol{\xi}$ when choosing $\Gamma = 1.0, 1.1$, or 1.2, offering one explanation for the strong performance of the allocations designed under these schemes. In a more general sense, the robust allocations under Γ serve as a form of regularization, bringing the design closer to an equal allocation to hedge against the possibility that $\hat{\tau}_o$ is far off from τ. In doing so, the robust designs may improve empirical performance even when the Tan model does not accurately characterize the form of unmeasured confounding.

5. Application to the Women's Health Initiative data

5.1. Setup

The Women's Health Initiative (WHI), a 1991 study of the effects of hormone therapy on health outcomes for postmenopausal women, comprises both an experimental arm, with 16,608 women enrolled in the trial, and an observational dataset of 53,054 women deemed clinically comparable to women in the trial. The treatment of interest was a 625 mg daily dosage of estrogen and a 2.5 mg dosage of progestin. Half the women in the RCT were randomized to receive the treatment, while about a third of women in the observational study were taking estrogen and progestin as part of a standard medical regimen [32]. In the WHI results, there was substantial disagreement in the causal effect estimates between the observational and experimental components, leading to several detailed reanalyses (see e.g. [19]).

Here, we ignore the realized WHI experimental results and pretend we are designing a small RCT of $n_r = 1{,}000$ units, with the intent of using $\boldsymbol{\kappa}_1$ to shrink the results toward those of the WHI observational study. Though many clinical outcomes were measured in the WHI, our interest is in the effect of hormone therapy on coronary heart disease (CHD). We plan to stratify our experiment on two clinically relevant variables: age and history of cardiovascular disease. The trial protocol discusses age as an important subgroup variable to consider [12], while subsequent papers note the importance of a history of cardiovascular disease [33].

The age variable has three levels: whether a woman was in her fifties, sixties, or seventies at the start of the trial. The history of cardiovascular disease history variable is a simple "yes" vs. "no" binary variable. The distributions of these variables can be found in Tables 5 and 6 in the Supplementary Material. We stratify on both variables, yielding a total of six strata.

5.2. Guardrails

We consider five experimental designs: a Neyman allocation, a "naïve" allocation (assuming $\boldsymbol{\xi} = 0$), and three robust allocations assuming worst-case error under $\Gamma = 1.0, 1.5$, and 2.0 respectively. As in the simulations, allocations are solved by running gradient descent (Algorithm 6) until we hit a stopping criterion, and then running greedy swapping (Algorithm 5) to convergence. Due to the non-convexity of the objective function for the naïve and robust designs, we again tried seven starting points: the equal allocation, the Neyman allocation, and five randomly chosen points. As in the simulations, we always reached the same final allocation of units to strata and treatment assignment, regardless of starting point.

We considered the guardrails discussed in Section 3.5. To preserve the viability of our Central Limit Theorem, we imposed a minimum sample size constraint of 30 units per stratum and treatment arm. To assess detachability, we evaluated the final designs relative to an equal allocation. We first evaluated detachability under $\hat{\boldsymbol{V}}$, the "direct" stratum potential outcome variance estimates obtained from the observational study, after a propensity score adjustment. As a robustness check, we also considered the three sets of stratum potential outcome variances associated with the worst case bias under each of our sensitivity models: $\tilde{\boldsymbol{V}}_{\mathbf{1.0}}, \tilde{\boldsymbol{V}}_{\mathbf{1.5}}$, and $\tilde{\boldsymbol{V}}_{\mathbf{2.0}}$. Reasonable performance under *all* of these candidate variance estimates increases our confidence that detachability will hold in the future experiment.

In Table 3, we report the value of δ_d—the ratio of the estimated risk under a given design to the estimated risk under equal allocation—for each possible method of computing the variance. Most values are less than 1.0, indicating that the design would be more efficient than an equal allocation if $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ were used on its own.

The largest value, which can be found in the first column in the final row, is only 1.02. This means that—were we to design our experiment using a robust

TABLE 3

Values for δ_d, the risk ratio using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, relative to an equal allocation, for each design (rows), computed under different estimates for the stratum potential outcome variances (columns).

	Variance Estimate			
Design	**Direct**	**$\Gamma = 1.0$**	**$\Gamma = 1.5$**	**$\Gamma = 2.0$**
Neyman	0.87	0.86	0.87	0.89
Naïve	0.89	0.88	0.90	0.93
$\Gamma = 1.0$	0.92	0.86	0.84	0.86
$\Gamma = 1.5$	0.96	0.86	0.80	0.78
$\Gamma = 2.0$	1.02	0.89	0.80	0.77

TABLE 4

Checking Condition 2 (the condition under which $\boldsymbol{\kappa}_1$ is guaranteed to have risk lower than $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$) for different designs and estimates of the potential outcome variances. A checkmark signifies that the condition holds and a dash signifies that the condition does not hold.

	Variance Estimate			
Design	**Direct**	**$\Gamma = 1.0$**	**$\Gamma = 1.5$**	**$\Gamma = 2.0$**
Neyman	–	–	–	–
Naïve	✓	✓	✓	✓
$\Gamma = 1.0$	✓	✓	✓	✓
$\Gamma = 1.5$.	✓	–	–	–
$\Gamma = 2.0$	–	–	–	–

allocation under $\Gamma = 2.0$ when the true stratum potential outcome variances were the "direct" estimates obtained from the observational study, and were we to use $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ alone as our estimator—we could incur a 2% risk penalty relative to using an equally allocated experiment. We consider this well within a reasonable tolerance range, given the potential upside to using $\boldsymbol{\kappa}_1$ instead of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$.

Lastly, we consider the risk reduction constraint. As with detachability, we consider the value of the constraint for each of our designs, under different estimates of the potential outcome variances. In Table 4, we show the results, with a checkmark signifying that the condition holds and a dash meaning it does not. Using the direct estimates of the potential outcome variances, $\hat{\boldsymbol{V}}$, we find that the naïve design and the robust designs under $\Gamma = 1.0$ and $\Gamma = 1.5$ yield the guarantee that that $\boldsymbol{\kappa}_{1+}$ dominates $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$. By contrast, under the three robust variance estimates—$\tilde{\boldsymbol{V}}_{\mathbf{1.0}}$, $\tilde{\boldsymbol{V}}_{\mathbf{1.5}}$, and $\tilde{\boldsymbol{V}}_{\mathbf{1.0}}$—only the naïve and $\Gamma = 1.0$ designs satisfy the risk reduction condition.

The relative sparsity of Table 4 might give us pause. If we believe the variance estimates under $\Gamma = 1.0, \Gamma = 1.5$ or $\Gamma = 2.0$ are plausible, for example, then we might want to recompute the designs incorporating Condition 2 as a constraint. However, we are afforded a certain amount of grace by the fact that we will be able to obtain unbiased estimates of the stratum potential outcome variances after the experiment is complete. Hence, we could instead proceed with one of these designs and defer the decision as to whether to use $\boldsymbol{\kappa}_{1+}$ instead of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ to a later date. For now, we will proceed with the unconstrained designs.

5.3. Allocation results

The allocations under each of our design heuristics are summarized in Figure 1. In the left panel, we provide the direct estimates of the stratum potential outcome variances, $\hat{\boldsymbol{V}}$, obtained from the WHI observational study. These estimates are corrected using stabilized inverse probability weighting, but may still be biased due to unmeasured confounding (for more details about the propensity score computation, see [37]).

There is a high level of variability in the stratum potential outcome variance estimates across strata. This follows from the fact that coronary heart disease is a relatively rare, binary outcome. Women in their fifties with no history of cardiovascular disease are very unlikely to have a coronary heart disease incident, while the incidence rate is much higher among older women or women with a history of cardiovascular disease. Because the variance of a binary outcome is a direct function of its mean, the corresponding potential outcome variance estimates fluctuate across strata. Note that the plotted values are the potential outcome variance estimates used as the input $\hat{\boldsymbol{V}}$ to the design algorithm for the Neyman and naïve allocations. The robust allocations, by contrast, use the variance estimates $\tilde{\boldsymbol{V}}_{\mathbf{1.0}}$, $\tilde{\boldsymbol{V}}_{\mathbf{1.5}}$, and $\tilde{\boldsymbol{V}}_{\mathbf{1.0}}$ corresponding to the worst-case bias under these models. Nonetheless, the broad trends in variability across strata are similar across the different potential outcome variance estimation methods. More detail can be found in Figure 2 in the Supplementary Material.

We make a number of observations about the allocations given in the right panel of Figure 1. The primary driver of all of the allocations is the estimated potential outcome variances in each stratum: each allocation oversamples high-variance strata and treatment arms, while undersampling low-variance strata and treatment arms. The algorithm likely would have allocated even fewer units to the stratum corresponding to women in their fifties with no history of cardiovascular disease, were a minimum threshold of 30 units not provided as a constraint.

In moving from a Neyman to a naïve allocation, we observe that this over- and undersampling behavior becomes more extreme. The naïve design samples more heavily from high variance strata and treatment arms, and less heavily from low variance strata and treatment arms. Recall that the naïve allocation assumes $\boldsymbol{\xi} = \mathbf{0}$. The sampling behavior may, then, reflect a notion that the observational study estimates are generally reliable, and resources are best allocated to obtaining precise experimental estimates in the highest variance strata and treatment arms, while an imprecise estimate in the low-variance strata and treatment arms can be sufficiently improved by shrinkage toward the observational estimate.

Interestingly, as we move to a robust allocation and progressively increase Γ from 1.0 to 1.5 to 2.0, this pattern begins to reverse. At $\Gamma = 1.0$, we see behavior similar to the naïve allocation, with perhaps slightly more aggressive oversampling in high variance strata and treatment arms. However, as Γ grows, the allocations are regularized back toward a more equal distribution across strata and treatment arms. This is likely due to the fact that, at larger values of Γ, the dominant form of error in the observational study estimates is bias,

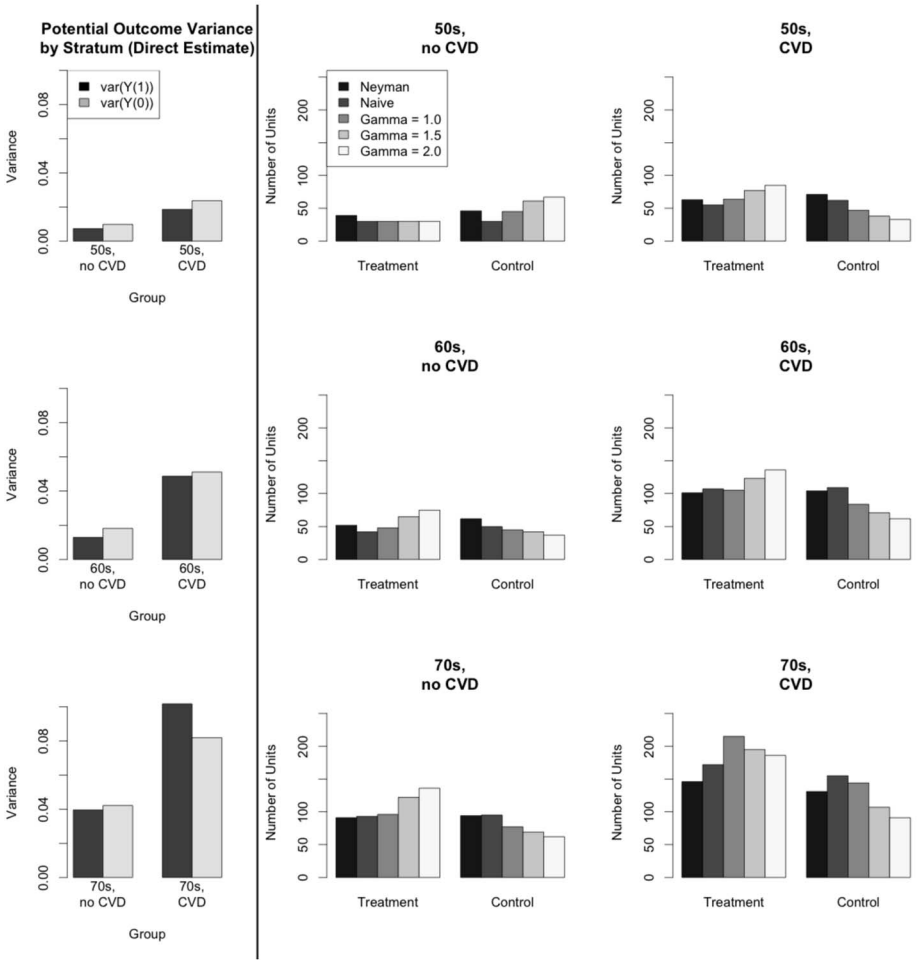

FIG 1. *Experimental allocation results for* $n_r = 1{,}000$. *The left panel shows the "direct" stratum potential outcome variance estimates* $\hat{\boldsymbol{V}} = \{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^{6}$ *obtained from the observational data. The right panel shows the allocations to each stratum and treatment arm under each prospective design.*

rather than variance. The observational study estimates cannot be relied upon in any of the strata, and it becomes sensible to obtain high-quality estimates from the RCT for all stratum treatment effects.

6. Discussion

We have considered the problem of designing a stratified experiment when we plan to use an Empirical Bayes estimator, such as $\boldsymbol{\kappa}_1$ or $\boldsymbol{\kappa}_{1+}$, to shrink the stratum causal estimates of our RCT toward estimates previously obtained from

an observational study. As we have shown, the risk of $\boldsymbol{\kappa}_1$ can be computed explicitly if the stratum potential outcome variances and the stratum-specific errors of the observational study estimates, $\boldsymbol{\xi}$, are known. In the absence of such information, we have proposed three heuristics—Neyman allocation, "naïve" allocation assuming $\boldsymbol{\xi} = 0$, and robust allocation under the worst-case value of $\boldsymbol{\xi}$ given a sensitivity model—for designing the experiment. We have also emphasized "detachability," the ability to do good causal inference using the RCT on its own, and suggested imposing constraints on the experimental design to ensure detachability is preserved.

We simulated a realistic scenario in which we have access to a large observational database and are interested in a relatively rare outcome. We considered a stratification comprising twelve strata, across which the outcome frequency varies dramatically. In this setting, there were significant risk improvements to be realized by using a shrinkage estimator, even if the design were a simple equal allocation. Additional gains were possible when designing the experiment explicitly for use with a shrinkage estimator. In simulations with and without unmeasured confounding in the observational study, we saw that the naïve allocation typically performed well, exhibiting surprising robustness to the case when $\boldsymbol{\xi}$ was far from zero. Our robust allocations performed well in the case when unmeasured confounding was present in the data.

Using data from the Women's Health Initiative, we designed a hypothetical 1,000-subject experiment that would incorporate these data. We explored a menu of possible designs for the experiment, including three possible allocations under different sensitivity models. We confirmed that the designs would all be reasonably "detachable" from the observational study, meaning that $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under these designs would incur statistical risk not much greater than an equally allocated RCT, under any of our plausible variance estimates. The designs also exhibited a notion of regularization: the Neyman, naïve and $\Gamma = 1.0$ designs relied increasingly heavily on the observational study in low-variance strata and treatment arms, allocating most of the RCT units instead to high-variance strata and treatment arms. As Γ rose to 1.5 and then 2.0, the algorithm contended with the possibility that $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ incorporates large biases in all of its strata estimates, and regularized the allocation back toward an equal distribution across strata and treatment arms.

A natural question is how our framework works for designing experiments using alternate definitions of outcomes, such as log odds ratios. The design ideas in principle extend, as estimators of these quantities tend to be asymptotically normal. To assess performance with non-normal outcomes we conducted additional simulations, following the template of our primary simulation. See the Supplementary Material for full details on these simulations. We find that indeed, our methods are valid when estimating quantities that are not simple averages. That being said, because asymptotic approximations can be slow to kick in, some gains in smaller sample contexts were negligible. In practice, we would encourage analysts to, as part of the planning of an experiment, simulate under expected stratum-specific parameter values to ensure that delta method variance approximations are valid given the planned size of the experiment.

There are many plausible extensions to this line of work. Approaches discussed in this paper have somewhat limited utility in smaller trials, which may be powered to estimate only an average treatment effect (ATE) rather than a vector of conditional average treatment effects. Future work should quantify the benefits for estimation of the overall ATE by weighting the stratum-specific estimates.

Another important use case is that of multiple observational studies. Interest in an experiment may be particularly high when several observational studies yield conflicting results. In this case, we may want methods that can incorporate more than one observational dataset. A simple approach is to pre-aggregate the evidence from the competing observational studies: one could simply pool the data together, or use a more complex weighting scheme to account for data quality and representativeness. Then, one could proceed with the existing techniques, extracting the estimates $\hat{\boldsymbol{\tau}}_{\boldsymbol{O}}$ and the pilot potential outcome variances from the merged observational data. A more complex approach would involve designing an estimator to incorporate multiple observational point estimates. The URE-minimization procedure described in [36] is a general purpose "recipe" for shrinkers, and makes it straightforward to design such an estimator. With such an estimator in hand, the design problem would be substantively similar to the one described in this manuscript: compute the unbiased risk estimate; treat the data from the multiple observational datasets as fixed; and estimate the risk via numerical integration for any desired heuristic discussed in Section 3.4.

On a similar note, adaptive procedures that incorporate the RCT data in waves—rather than all at once—have direct utility when combined with our design procedure. The veracity of Assumption 2 and Assumption 3 cannot be ascertained until RCT data is collected, and strong subject matter knowledge is required to assess their plausibility. Adaptive approaches would allow researchers to collect some data to assess the validity of these assumptions and, optionally, revert to a more standard design in later data-collection waves if the assumptions do not hold. Such an approach is analogous to pre-test estimation in the Empirical Bayes literature [see e.g. 39, 24]. The test itself could be conducted via a standard F-testing approach or using more modern methods from the causal inference literature [47]. This is a promising future direction for this line of research.

Lastly, we have assumed a rigid model for treatment effect heterogeneity in this paper: treatment effects vary according to a known stratification on observed covariates. More modern work [25, 28] focuses on estimating heterogeneous treatment effects empirically, allowing for greater flexibility in how the effects differ across units. Incorporating such ideas into this work, we might imagine using some of the observational data in a first stage to estimate a stratification scheme, and using the remaining data for shrinkage in a second stage. Alternatively, we could consider modeling the causal effect explicitly as a function of the covariates, e.g. defining $\tau(\boldsymbol{X}_i)$ rather than a vector of true treatment effects $\boldsymbol{\tau}$. We could then define a flexible shrinker to trade off between estimates not within strata, but within nearby values of the covariates themselves.

Supplementary Material

Alternative shrinkers to $\boldsymbol{\kappa_1}$

We have supposed that, after the experiment's conclusion, we will use $\boldsymbol{\kappa}_1$ or $\boldsymbol{\kappa}_{1+}$ to shrink causal estimates from $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ toward $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$. However, our results are not dependent on the form of the shrinkage estimator. Our goal is to apply Theorem 3.1 in [41] to compute the exact risk of the estimator. Hence, we can use any estimator $\boldsymbol{\theta}$ of the form

$$\boldsymbol{\theta} = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} + \boldsymbol{\Sigma}_r g_\theta(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})$$

where $g_\theta(x, y)$ is weakly differentiable in x and $\mathbb{E}_r\left(\|g_\theta\|^2\right) < \infty$.

A plausible alternative estimator proposed in [36] is

$$\boldsymbol{\kappa}_2 = \hat{\boldsymbol{\tau}}_{\boldsymbol{o}} + \left(\boldsymbol{I}_K - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r^2)\boldsymbol{\Sigma}_r}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^\mathsf{T}\boldsymbol{\Sigma}_r^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})}\right)(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})\,.$$

and its positive part analogue, $\boldsymbol{\kappa}_{2+}$. Two other standard alternatives, introduced in [16], are

$$\boldsymbol{\delta}_1 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} + \left(\frac{K-2}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^\mathsf{T}\boldsymbol{\Sigma}_r^{-1}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}\right)(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})$$

and

$$\boldsymbol{\delta}_2 = \hat{\boldsymbol{\tau}}_{\boldsymbol{r}} + \left(\frac{(K-2)\boldsymbol{\Sigma}_r^{-1}}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^\mathsf{T}\boldsymbol{\Sigma}_r^{-2}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}\right)(\hat{\boldsymbol{\tau}}_{\boldsymbol{o}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}).$$

Each also has a positive part version that is straightforward to define.

Once a suitable estimator $\boldsymbol{\theta}$ is chosen, we can redefine our procedures to work with that estimator. The first step is to apply Theorem 3.1 from [41] to compute its risk conditional on $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$,

$$\mathcal{R}(\theta) = \frac{1}{K}\left(\operatorname{tr}(\boldsymbol{\Sigma}_r) + \mathbb{E}_r\left(\sum_k \sigma_{rk}^4\left(g_{\theta,k}^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}) + 2\frac{\partial g_{\theta,k}(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}{\partial \tau_{rk}}\right)\right)\right).$$

Next, we can deploy the method from Section 3.2 to obtain an exact integral expression for the the risk of the estimator. [3] contains a detailed explanation of how to construct each component of the integral. For example, if we use $\boldsymbol{\kappa}_2$, we obtain

$$\mathcal{R}(\boldsymbol{\kappa}_2) = \frac{1}{K}\Bigg(\operatorname{tr}(\boldsymbol{\Sigma}_r) + \operatorname{tr}(\boldsymbol{\Sigma}_r^2)\mathbb{E}_r\left(\frac{4(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^\mathsf{T}\boldsymbol{\Sigma}_r^4(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}{((\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^\mathsf{T}\boldsymbol{\Sigma}_r^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}}))^2} - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_r^2)}{(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})^\mathsf{T}\boldsymbol{\Sigma}_r^2(\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o}})}\right)\Bigg),$$

and the relevant integrals can be computed as

$$\begin{aligned}
\mathbb{E}_r\left(\frac{\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^5\boldsymbol{\nu}}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^3\boldsymbol{\nu})^2}\right) &= \int_0^\infty \det(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r^3)^{-1/2} \cdot \\
&\quad \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1/2}(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\xi}\right)\right) \\
&\quad \left(\mathrm{tr}(\boldsymbol{R}) + (\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})^\mathsf{T}\boldsymbol{R}(\boldsymbol{L}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi})\right) t dt \\
\mathbb{E}_r\left(\frac{1}{(\boldsymbol{\nu}^\mathsf{T}\boldsymbol{\Sigma}_r^3\boldsymbol{\nu})}\right) &= \int_0^\infty \det(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r^3)^{-1/2} \cdot \\
&\quad \exp\left(\frac{1}{2}\left(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1/2}(\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r)^{-1}\boldsymbol{\Sigma}_r^{-1/2}\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\xi}\right)\right) dt
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
\boldsymbol{L} &= (\boldsymbol{I}_K + 2t\boldsymbol{\Sigma}_r^3)^{-1/2}, \qquad \text{and} \\
\boldsymbol{R} &= \boldsymbol{L}^\mathsf{T}\boldsymbol{\Sigma}_r^5\boldsymbol{L}.
\end{aligned}$$

This process can be repeated for any estimator $\boldsymbol{\theta}$. Once integral expressions for the risk have been obtained, the design heuristics described in Section 3.4 can be deployed, using $\mathcal{R}(\boldsymbol{\theta})$ as the objective function rather than $\mathcal{R}(\boldsymbol{\kappa}_1)$.

Robust versions of constraints

A more robust version of the detachability constraint can be imposed if the analyst assumes the Tan sensitivity model, as discussed in Section 3.4.3. Under a given choice of Γ and α, we can obtain bounds on the potential outcome means in each stratum. We reject any design such that

$$\max_{\substack{\mu_{kc}\in(\ell_{kc}^{(\Gamma,\alpha)},u_{kc}^{(\Gamma,\alpha)}),\\ \mu_{kt}\in(\ell_{kt}^{(\Gamma,\alpha)},u_{kt}^{(\Gamma,\alpha)})}} \frac{\sum_k \frac{\mu_{kt}(1-\mu_{kt})}{n'_{rkt}} + \frac{\mu_{kt}(1-\mu_{kt})}{\tilde{n}_{rkc}}}{\sum_k \frac{\mu_{kt}(1-\mu_{kt})}{n'_{rkt}} + \frac{\mu_{kc}(1-\mu_{kc})}{\tilde{n}_{rkc}}} \geq \delta_d. \tag{9}$$

In words, this means that we are rejecting any design $\boldsymbol{d'}$ such that the risk of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ under design d' is larger than the estimated risk under the default design by a factor greater than δ_d, for *any* configuration of the potential outcome means consistent with our sensitivity bounds. Practically, the left-hand-side of Inequality (9) can be reduced to a quadratic fractional programming problem and solved via Dinkelbach's method [10].

The robust version of the risk reduction incorporates the parameter bounds from the Tan sensitivity model. We reject any design such that

$$\begin{aligned}
4\max_k \min_{\substack{\mu_{kc}\in(\ell_{kc}^{(\Gamma,\alpha)},u_{kc}^{(\Gamma,\alpha)}),\\ \mu_{kt}\in(\ell_{kt}^{(\Gamma,\alpha)},u_{kt}^{(\Gamma,\alpha)})}} &\left(\frac{\mu_{kt}(1-\mu_{kt})}{n'_{rkt}} + \frac{\mu_{kc}(1-\mu_{kc})}{n'_{rkc}}\right)^2 > \\
&\max_{\substack{\mu_{kc}\in(\ell_{kc}^{(\Gamma,\alpha)},u_{kc}^{(\Gamma,\alpha)}),\\ \mu_{kt}\in(\ell_{kt}^{(\Gamma,\alpha)},u_{kt}^{(\Gamma,\alpha)})}} \sum_k \left(\frac{\mu_{kt}(1-\mu_{kt})}{n'_{rkt}} + \frac{\mu_{kc}(1-\mu_{kc})}{n'_{rkc}}\right)^2.
\end{aligned}$$

WHI experimental design: further details

The age variable has three levels: whether a woman was in her fifties, sixties, or seventies at the start of the trial. The history of cardiovascular disease history variable is a simple "yes" vs. "no" binary variable. The distributions of these variables can be found in Tables 5 and 6 below.

TABLE 5
Distribution of age variable values in the WHI observational study and RCT.

Age	Observational Study	RCT
50–59	17,447 (33.0%)	5,491 (33.2%)
60–69	23,030 (43.6%)	7,473 (45.2%)
70–79	12,388 (23.4%)	3,573 (21.2%)

TABLE 6
Distribution of history of cardiovascular disease in the WHI observational study and RCT.

History of CVD	Observational Study	RCT
Yes	8,709 (16.5%)	1,828 (11.1%)
No	44,156 (83.5%)	14,709 (88.9%)

In Figure 2, we provide a slightly more detailed version of Figure 1, where we include the variance estimates under each of the sensitivity models in the left panel of the plot. Observe that the Neyman and naïve allocations both utilize the direct variance estimates, so those are plotted a single time but with double the width in the left panel.

Extension to multiple observational studies

There are two potential avenues to extend this approach to the "multiple observational studies" use case: a simple approach based on pre-aggregation, and a more complex approach that requires use of a different estimator.

We calculate a design using two summary elements of the observational study: a vector of stratum-specific causal estimates $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ and a matrix of stratum-specific potential outcome variance estimates $\{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K$. In the case of multiple observational studies, our first general approach is to aggregate the studies to generate cross-study estimates of $\hat{\boldsymbol{\tau}}_{\boldsymbol{o}}$ and $\{(\hat{\sigma}^2_{kt}, \hat{\sigma}^2_{kc})\}_{k=1}^K$, and then proceed as though we only had a single observational study. A direct way for doing this, if we had the raw data, would be to simply pool the observational data and estimate on the full dataset. We could instead leverage meta-analysis techniques to aggregate the estimates rather than the data, e.g. we could weight the impact vector and variance estimates based on data quality.

Alternatively, we could construct a shrinkage estimator designed specifically to incorporate multiple observational studies. The methods discussed in [36] can be readily adapted to develop such a shrinker. We work through an example below where there are two observational studies (rather than one), and we want to utilize a shrinker that uses the same weighting scheme across all strata (analogous to $\boldsymbol{\kappa}_1$ in the single-observational-study case). However, these methods are

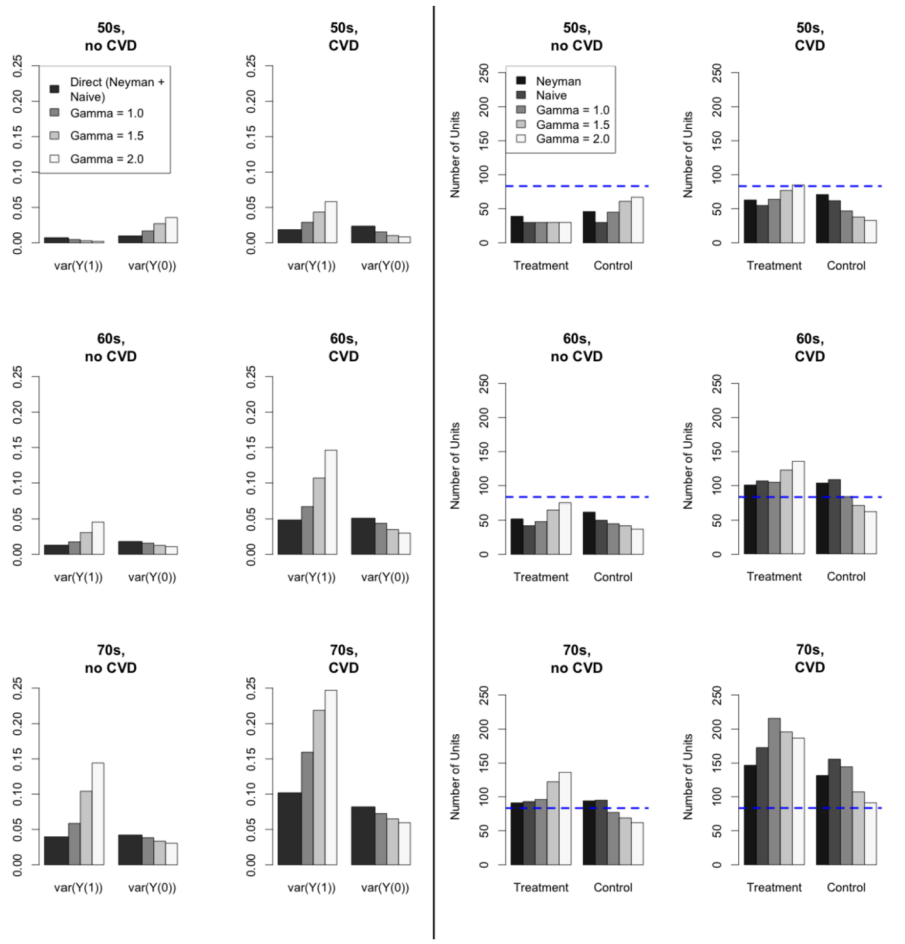

FIG 2. *Experimental allocation results for* $n_r = 1{,}000$. *The left panel shows the direct stratum potential outcome variance estimates* $\hat{\boldsymbol{V}}$ *as well as the estimates* $\hat{\boldsymbol{V}}_{\mathbf{1.0}}, \hat{\boldsymbol{V}}_{\mathbf{1.5}}$, *and* $\hat{\boldsymbol{V}}_{\mathbf{2.0}}$ *under each of the sensitivity models. The right panel shows the allocations to each stratum and treatment arm under each prospective design. The reference line in blue in the right panel reflects an equal allocation across all strata and treatment levels.*

quite flexible and could incorporate more observational studies or alternative shrinker constructions.

Denote as $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ the vector of estimates from the RCT, and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}}$ the vector of estimates from the two observational studies. Utilizing the method in [36] we can use URE minimization to obtain the shrinker:

$$\boldsymbol{\psi} = (1 - \lambda_1 - \lambda_2)\hat{\boldsymbol{\tau}}_{\boldsymbol{r}} + \lambda_1 \hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} + \lambda_2 \hat{\boldsymbol{\tau}}_{\boldsymbol{o2}}$$

where

$$\lambda_1 = \max\left(\frac{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}(\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o1}})\,\mathrm{tr}(\Sigma_r)}{2(\|\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\|_2^2 \|\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\|_2^2 - ((\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}(\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}))^2)}, 0\right) \quad \text{and}$$

$$\lambda_2 = \max\left(\frac{(\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}(\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{o2}})\,\mathrm{tr}(\Sigma_r)}{2(\|\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\|_2^2\|\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}\|_2^2 - ((\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}})^{\mathsf{T}}(\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}} - \hat{\boldsymbol{\tau}}_{\boldsymbol{r}}))^2)}, 0\right).$$

Using this estimator, the design problem would be substantively similar: compute the unbiased risk estimate using the expression given in [36]; condition on the values of $\hat{\boldsymbol{\tau}}_{\boldsymbol{o1}}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{o2}}$; and estimate the risk via numerical integration under the three different heuristics discussed in the paper.

The only remaining question would be how to obtain pilot estimates of the stratum-specific potential outcome variance estimates $\{(\hat{\sigma}_{kt}^2, \hat{\sigma}_{kc}^2)\}_{k=1}^K$. As in the pre-aggregation approach, one could simply pool the data to estimate these variances. Alternatively, one could utilize a meta-analysis technique to account for study quality, or any other data-combination approach.

Robustness to non-normality of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$

In this Supplementary Material, we explore how well our framework operates in non-normal contexts. Generally, our framework should be applicable to experiments measuring such quantities as log odds or hazard ratios, as these estimators tend to also be asymptotically normal. That being said, we find gains to be muted in some contexts, as we discuss next.

First, the simulations in Section 4 indicate that the main results are fairly robust to at least some types of non-normality. Recall that these simulations involve 1,000 units across 12 strata and two treatment statuses, and the incidence rate of the outcome is, on average, 10% in the treatment group and 11% in the control group. Hence, even in an equally allocated experiment, most strata would fail to satisfy the so-called "success-failure" condition stating that normality can be assumed for an average of binary outcomes only if there are at least 10 expected "successes" (e.g. outcomes of $Y_i = 1$) and 10 expected "failures" (e.g. outcomes $Y_i = 0$) per draw. In our case, for many of the Neyman, naïve, and robust designs, upwards of 50% of the strata k are expected to have fewer than three successes due to the rarity of the outcome. Even so, the simulation results broadly indicate that we are able to achieve gains by applying a descent algorithm on the estimated risk of the shrinker. This suggests reasonably good robustness to non-normality.

The optimization depends on the assumed potential outcome variances of each stratum in the experiment. Section 3.3 discusses the assumptions that are necessary for estimating these quantities when using a difference-in-means estimator. In the case of non-linear functions of the outcomes – such as log odds ratios – there is another complication: if the number of units is small, the delta method variance approximations may not be valid, and hence yield incorrect estimates of $\boldsymbol{\Sigma}_r$. This is true even if Assumption 3 and unconfoundedness hold exactly.

For small experiments, the delta method variance estimates for quantities such as log odds ratios are often not particularly accurate. We investigated this issue by rerunning the simulations from Section 4 of the paper, but seeking to estimate the log odds ratio rather than the difference in means. For simplicity,

TABLE 7

Relative risk over 25,000 *iterations of* $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}, \boldsymbol{\kappa}_1$, *and* $\boldsymbol{\kappa}_{1+}$, *estimating a log odds ratio with a sample size of 1,000 units, under various experimental designs. Designs under the header "Estimated Parameters" use the observational study to estimate the stratum-specific potential outcome means, whereas those under "True Parameters" are computed under the true stratum-specific potential outcome means.*

			Estimated Parameters			**True Parameters**	
Est.	**Unmeasured Confounding?**	**Trt.**	**Equal**	**Neyman**	**Naïve**	**Neyman**	**Oracle**
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	118%	126%	114%	121%
$\boldsymbol{\kappa}_1$	no	c	116%	91%	89%	84%	82%
$\boldsymbol{\kappa}_{1+}$			57%	55%	55%	52%	53%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	113%	120%	112%	121%
$\boldsymbol{\kappa}_1$	no	ℓ	84%	67%	66%	61%	60%
$\boldsymbol{\kappa}_{1+}$			19%	19%	19%	19%	19%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	123%	132%	124%	132%
$\boldsymbol{\kappa}_1$	no	q	132%	93%	88%	92%	88%
$\boldsymbol{\kappa}_{1+}$			24%	24%	24%	24%	24%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	116%	122%	116%	119%
$\boldsymbol{\kappa}_1$	yes	c	85%	84%	86%	81%	83%
$\boldsymbol{\kappa}_{1+}$			63%	66%	65%	63%	64%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	105%	108%	109%	109%
$\boldsymbol{\kappa}_1$	yes	ℓ	72%	80%	74%	77%	73%
$\boldsymbol{\kappa}_{1+}$			61%	67%	64%	66%	64%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	123%	124%	134%	130%
$\boldsymbol{\kappa}_1$	yes	q	120%	132%	117%	124%	117%
$\boldsymbol{\kappa}_{1+}$			92%	101%	92%	99%	93%

we considered only the equal, Neyman, and naïve designs. Results for both simulation settings given in the paper (no unmeasured confounding, and unmeasured confounding induced by omitting the third variable) are given below in Table 7. We again suppose a sample size of 1,000 units in the RCT and 20,000 units in the observational study. For variance computations, we used the standard delta method approximation to the variance of log odds ratios (given by the sum of the reciprocals of the table-cell-specific counts).

Unlike the simulations estimating a difference in means, we do *not* find that the Neyman and naïve designs consistently outperform the equal allocation design across all choices of estimator. These designs perform better for the shrinker $\boldsymbol{\kappa}_1$ and its positive-part analogue $\boldsymbol{\kappa}_{1+}$ (as intended), but the equal design now performs better when we use $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$ on its own.

We then investigated why the Neyman design failed to yield superior results when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, given that the Neyman allocation does not rely on normality. Our answer lies in the final two columns of the table, that show two additional designs where we conduct an "oracle" optimization that has access to the true stratum-specific potential outcome means and observational study bias parameters. The first design is a Neyman allocation under the true parameters; the second is the "oracle" design provided in the manuscript, where we optimize the risk of $\boldsymbol{\kappa}_1$ under the true parameter values. We observe that the Neyman allocation *with access to the true stratum-specific potential outcome means* still does not outperform the equal allocation design when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$. This is because the delta

TABLE 8

Risk over 25,000 iterations of $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, $\boldsymbol{\kappa}_1$, and $\boldsymbol{\kappa}_{1+}$, estimating a log odds ratio with a sample size of 10,000 units, under various experimental designs. Designs under the header "Estimated Parameters" use the observational study to estimate the stratum-specific potential outcome means, whereas those under "True Parameters" are computed under the true stratum-specific potential outcome means.

			Estimated Parameters			True Parameters	
Est.	**Unmeasured Confounding?**	**Trt.**	**Equal**	**Neyman**	**Naïve**	**Neyman**	**Oracle**
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	79%	82%	78%	79%
$\boldsymbol{\kappa}_1$	no	c	55%	44%	43%	43%	41%
$\boldsymbol{\kappa}_{1+}$			48%	41%	41%	41%	40%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	78%	85%	80%	81%
$\boldsymbol{\kappa}_1$	no	ℓ	65%	53%	54%	53%	53%
$\boldsymbol{\kappa}_{1+}$			62%	53%	54%	52%	53%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	81%	86%	78%	78%
$\boldsymbol{\kappa}_1$	no	q	78%	66%	70%	65%	65%
$\boldsymbol{\kappa}_{1+}$			78%	66%	70%	65%	65%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	84%	84%	79%	76%
$\boldsymbol{\kappa}_1$	yes	c	83%	71%	71%	67%	66%
$\boldsymbol{\kappa}_{1+}$			83%	71%	71%	67%	66%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	90%	89%	81%	82%
$\boldsymbol{\kappa}_1$	yes	ℓ	86%	78%	77%	71%	72%
$\boldsymbol{\kappa}_{1+}$			86%	78%	77%	71%	72%
$\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$			100%	83%	83%	76%	76%
$\boldsymbol{\kappa}_1$	yes	q	86%	74%	74%	69%	68%
$\boldsymbol{\kappa}_{1+}$			86%	74%	74%	69%	68%

method variance computation is not accurate in this regime, so even knowing the true stratum-specific potential outcome means does not give us a good estimate of the estimator variance.

In Table 8, we rerun the simulations with a much larger RCT sample size of 10,000 (leaving the size of the observational study unchanged). In this regime, we see results more similar to those in the manuscript. Namely, the Neyman and naïve allocations always perform better than the equal allocation, regardless of the choice of estimator. In most cases, the Neyman allocation achieves slightly better performance when using $\hat{\boldsymbol{\tau}}_{\boldsymbol{r}}$, while the naïve allocation typically yields slightly better performance when using the shrinkage estimators. The allocations in the final two columns, computed using the true parameter values, also perform as expected. The asymptotic normality is holding in this larger sample size case.

Taken together, these simulations indicate that our methods are valid when estimating quantities that are not simple averages. However, additional caution should be taken when considering such quantities because asymptotic approximations may be slow to kick in. We would encourage analysts to simulate under expected stratum-specific parameter values to ensure that delta method variance approximations are valid given the expected size of the experiment.

References

[1] Athey, S., Chetty, R., Imbens, G. W. and 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.

[2] Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review* **112** 3911–3940.

[3] 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. MR3053545

[4] Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** 7345–7352. https://doi.org/10.1073/pnas.1510507113

[5] Chambaz, A., van der Laan, M. J. and Zheng, W. (2014). Targeted covariate-adjusted response-adaptive lasso-based randomized controlled trials. *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects* 345–368. MR3676364

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

[7] Cheng, D. and Cai, T. (2021). Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*.

[8] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

[9] Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J. and Yang, S. (In Press). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*.

[10] Dinkelbach, W. (1967). On nonlinear fractional programming. *Management Science* **13** 492–498. MR0242488

[11] Dorn, J. and Guo, K. (2022). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association* 1–13.

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

[13] Gagnon-Bartsch, J. A., Sales, A. C., Wu, E., Botelho, A. F., Erickson, J. A., Miratrix, L. W. and Heffernan, N. T. (2023). Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference* **11** 20220011. MR4631978

[14] Grant, M., Boyd, S. and Ye, Y. (2006). Disciplined convex programming. *Global Optimization: From Theory to Implementation* 155–210. MR2206954

[15] 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** 1001–1006. MR1146348
[16] 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** 1–6.
[17] FDA Draft Guidance (2018). Adaptive Designs for Clinical Trials of Drugs and Biologics. *Center for Biologics Evaluation and Research (CBER)*.
[18] Hahn, J., Hirano, K. and Karlan, D. (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics* **29** 96–108. MR2789394
[19] Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., Willett, W. C., Manson, J. E. and Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)* **19** 766.
[20] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20** 217–240. MR2816546
[21] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA. MR3309951
[22] Kallus, N., Puli, A. M. and Shalit, U. (2018). Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems* 10888–10897.
[23] Kallus, N. and Zhou, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science* **67** 2870–2890.
[24] Khan, S. and Md. Ehsanes Saleh, A. (1997). Shrinkage pre-test estimator of the intercept parameter for a regression model with multivariate Student-t errors. *Biometrical Journal* **39** 131–147. MR1453430
[25] Lee, K., Small, D. S. and Dominici, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *Journal of the American Statistical Association* **116** 569–580. MR4270004
[26] Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112** 1759–1769. MR3750897
[27] Neyman, J. (1992). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics* 123–150. Springer.
[28] Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133
[29] Oberst, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D. and Yadlowsky, S. (2023). Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467v2*.
[30] Offer-Westort, M., Coppock, A. and Green, D. P. (2021). Adap-

tive experimental design: Prospects and applications in political science. *American Journal of Political Science*.

[31] Peysakhovich, A. and Lada, A. (2016). Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385*.

[32] Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L. 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** 404–414.

[33] 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**.

[34] Rosenbaum, P. R. (2009). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612

[35] Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

[36] Rosenman, E. T. R., Basse, G., Owen, A. B. and Baiocchi, M. (2023). Combining observational and experimental datasets using shrinkage estimators. *Biometrics*. https://doi.org/10.1111/biom.13827

[37] Rosenman, E. T., Owen, A. B., Baiocchi, M. and Banack, H. R. (2022). Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine* **41** 65–86. MR4376789

[38] Rosenman, E. T. R. and Owen, A. B. (2021). Designing experiments informed by observational studies. *Journal of Causal Inference* **9** 147–171. https://doi.org/doi:10.1515/jci-2021-0010. MR4289528

[39] Sen, P. K. and Saleh, A. E. (1987). On preliminary test and shrinkage M-estimation in linear models. *The Annals of Statistics* 1580–1592. MR0913575

[40] Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 465–472. MR1092986

[41] Strawderman, W. E. (2003). On minimax estimation of a normal mean vector for general quadratic loss. In *Mathematical Statistics and Applications: Festschrift for Constance Van Eeden* 3–14. Institute of Mathematical Statistics. MR2138282

[42] Tabord-Meehan, M. (2023). Stratification trees for adaptive randomisation in randomised controlled trials. *Review of Economic Studies* **90** 2646–2673. MR4636650

[43] Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101** 1619–1637. MR2279484

[44] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**

285–294.

[45] Thompson, W. R. (1935). On the theory of apportionment. *American Journal of Mathematics* **57** 450–456. MR1507085

[46] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228–1242. MR3862353

[47] Yang, S., Gao, C., Zeng, D. and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85** 575–596.

[48] 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)* **81** 735–761. MR3997099