Optimal Estimation when Researcher and Social Preferences are Misaligned

Jann Spiess

Job Market Paper

Current version: `scholar.harvard.edu/spiess/jmp`

This version: December 8, 2017

Abstract

Econometric analysis typically focuses on the statistical properties of fixed estimators and ignores researcher choices. In this article, I approach the analysis of experimental data as a mechanism-design problem that acknowledges that researchers choose between estimators, sometimes based on the data and often according to their own preferences. Specifically, I focus on covariate adjustments, which can increase the precision of a treatment-effect estimate, but open the door to bias when researchers engage in specification searches. First, I establish that unbiasedness is a requirement on the estimation of the average treatment effect that aligns researchers' preferences with the minimization of the mean-squared error relative to the truth and is optimal in a minimax sense. Second, I provide a constructive characterization of all unbiased treatment-effect estimators as sample-splitting procedures. Third, I show that a researcher restricted to the class of unbiased estimators of the average treatment effect solves a prediction problem. The equivalence of unbiased estimation and prediction across sample splits characterizes all admissible procedures in finite samples, leaves space for beneficial specification searches, and offers an opportunity to leverage machine learning. As a practical implication, I describe flexible pre-analysis plans for randomized experiments that achieve efficiency without bias.

Jann Spiess, Department of Economics, Harvard University, `jspiess@fas.harvard.edu`. For their guidance I am indebted to Sendhil Mullainathan, Alberto Abadie, Elie Tamer, and Gary Chamberlain. For valuable comments or discussions I also thank Laura Blattner, Avi Feller, Edward Glaeser, Nathaniel Hendren, Simon Jäger, Maximilian Kasy, Lawrence Katz, Scott Kominers, Eben Lazarus, Shengwu Li, Mikkel Plagborg-Møller, Ashesh Rambachan, Jonathan Roth, Elizabeth Santorella, and seminar audiences at the Harvard Economics Department and the Harvard Center for Mathematical Sciences and Applications.

Introduction

There is a tension between flexibility and robustness in empirical work. Consider an investigator who estimates a treatment effect from experimental data. If the investigator has the freedom to choose a specification that adjusts for control variables, her choice can improve the precision of the estimate. However, the investigator's specification search may also produce an estimate that reflects a preference for publication or ideology instead of a more precise guess of the truth.[1] To solve this problem, we sometimes tie the investigator's hands and restrict her to a simple specification, like a difference in averages. In contrast, this article characterizes flexible estimators that leverage the data and researchers' expertise, and do not also reflect researchers' preferences.

To characterize optimal estimators when researcher and social preferences are misaligned, I approach the analysis of experimental data as a mechanism-design problem.[2] Concretely, I consider a designer and an investigator who are engaged in the estimation of an average treatment effect. As the designer, we aim to obtain a precise estimate of the truth (which I capture in terms of mean-squared error). I assume however that the investigator may care about the value of the estimate and not only its precision. For example, the investigator may have a preference for large estimates in order to get published. The investigator picks an estimator based on her private information about the specific experiment. The designer chooses optimal constraints on the estimation by the investigator.[3]

First, I establish that we should require that the investigator commits to an unbiased estimator. More precisely, I prove that fixing the bias is a minimax optimal solution to the designer's problem and aligns the incentives of the investigator and the designer under reasonable assumptions on preferences.[4] Allowing for bias can,

[1] A literature in statistics dating back to at least Sterling (1959) and Tullock (1959), and most strongly associated with the work of Edward Leamer (e.g. Leamer, 1974, 1978), acknowledges that empirical estimates reflect not just data, but also researcher motives. Fears of biases have been fueled more recently by replication failures (Open Science Collaboration, 2015), anomalies in published p-values (Brodeur et al., 2016), and empirical evidence for publication biases (Andrews and Kasy, 2017). This concern is also evident in the American Economic Association's 2012 decision to establish a registry for randomized controlled trials.

[2] Like Leamer (1974, 1978), I explicitly consider researchers' degrees of freedom. Like Glaeser (2006), I also model their preferences.

[3] Abstractly, the designer could represent professional norms. Concretely, it could represent a journal setting standards for the analysis of randomized controlled trials, or the U.S. Food and Drug Administration (FDA) imposing rules for the evaluation of new drugs.

[4] This result echoes Frankel's (2014) characterization of simple delegation mechanisms that align

in principle, improve overall precision through a reduction in the variance. But an investigator could use her control over the bias to reflect her preferences rather than her private information. Among unbiased estimators, however, even an investigator who wants to obtain an estimate close to some large, fixed value will still choose an estimator that minimizes the variance.

Second, having motivated a restriction to unbiasedness, I prove that every unbiased estimator of the average treatment effect has a sample-splitting representation. As the starting point for this representation, consider a familiar estimator that is unbiased, namely the difference in averages between treatment and control groups. We can adjust this estimator for control variables by a procedure that splits the sample into two groups. From the first group, we calculate regression adjustments that we subtract from the outcomes in the second group. The updated difference in averages is still unbiased by construction. Though this procedure appears specific, I prove that any unbiased estimator can be represented by multiple such sample-splitting steps. In particular, unbiased estimators can differ from a difference in averages only by leave-one-out or leave-two-out regression adjustments of individual outcomes.[5]

Third, I show that an investigator restricted to unbiasedness will solve a prediction problem. By the sample-splitting representation, I can write every unbiased estimator of the average treatment effect in terms of a set of regression adjustments. When choosing from this restricted set of estimators, the investigator picks regression adjustments that minimize prediction risk for a specific loss function. Each optimal adjustment predicts the outcomes of one or two units from other units in the sample.

The investigator's solution reveals a finite-sample complete-class theorem that characterizes all admissible unbiased estimators of an average treatment effect as solutions to out-of-sample prediction problems. Since my results hold exactly without taking large-sample limits or relying on other approximations, I obtain a general duality between unbiased estimation and prediction without putting any essential restrictions on the distribution of the data other than random assignment of treatment. Any admissible unbiased estimator corresponds exactly to a set of admissible

an agent's choices with a principal's preferences by fixing budgets. In Section 4, I explore the similarities of my solution to results in the mechanism-design literature on delegation that goes back to Holmström (1978, 1984), and I exploit these parallels in the proof of my minimax result.

[5]In particular, for known treatment probability, I show that all unbiased estimators of the sample-average treatment effect take the form of the "leave-one-out potential outcomes" (LOOP) estimator from Wu and Gagnon-Bartsch (2017), which is a special case of Aronow and Middleton's (2013) extension of the Horvitz and Thompson (1952) estimator.

prediction solutions.

As a practical implication, my results motivate and describe flexible yet robust pre-analysis plans for the analysis of experimental data.[6] Having established that unbiased estimation is equivalent to a set of prediction tasks, there are two types of flexible pre-analysis plan that achieve precise estimation of treatment effects without leaving room for bias from specification searches. In the first type, the investigator commits to an algorithm that predicts outcomes from covariates. This algorithm can engage in automated specification searches to learn a good model from the data.[7] Adjusting outcomes by its fitted out-of-sample predictions will yield an unbiased estimator.

There is a second, more flexible type of pre-analysis plan that achieves unbiased and precise estimation without the investigator committing to her specification searches in advance. In this second type of pre-analysis plan, the investigator only commits to splitting the data and distributing subsamples to her research team. Each researcher then engages in specification searches on a part of the data and reports back a prediction function. As my fourth main result, I characterize all unbiased estimators of the treatment effect that delegate the estimation of some or all regression adjustments in this way. Delegation to one researcher improves over simple pre-analysis plans.[8] Delegation to at least two researchers asymptotically attains the semi-parametric efficiency bound of Hahn (1998) under assumptions that apply to most parametric and many semi- and non-parametric estimators of the regression adjustments.

The results in this article relate to the practice of sample splitting in econometrics, statistics, and machine learning. From Hájek (1962) to Jackknife IV (Angrist et al., 1999), model selection (e.g. Hansen and Racine, 2012), and time-series forecasting (see e.g. Diebold, 2015; Hirano and Wright, 2017), sample splitting is used as a tool to avoid bias by construction. Wager and Athey (2017) highlight the role of sample splitting in the estimation of heterogeneous treatment effects. Chernozhukov

[6] Coffman and Niederle (2015), Olken (2015), and Heckman and Singer (2017) discuss the benefits, costs, and limitations of pre-analysis plans. I resolve an implicit flexibility-robustness tradeoff for one specific setting.

[7] In a similar spirit, Balzer et al. (2016) propose a data-adaptive procedure that selects among specifications to minimize the variance of treatment-effect estimators in experiments.

[8] My hold-out approach is similar to Dahl et al. (2008), Fafchamps and Labonne (2016) and Anderson and Magruder (2017), who all propose split-sample strategies to combine exploratory data analysis with valid inference. Dwork et al. (2015) propose a protocol to reuse the hold-out data to improve efficiency. I show that in my setting simple hold-out procedures are dominated when data can be distributed to multiple researchers.

et al. (2017b) show its relevance in achieving valid and efficient inference in high-dimensional observational data. My results show that sample splitting is not just an ad-hoc tool, but a feature of optimal estimators. I establish that sample splitting is a necessary restriction on the investigator's estimator to achieve unbiasedness and align incentives.

Moreover, this article contributes to a growing literature that employs machine learning in program evaluation. Supervised machine learning algorithms solve prediction problems like those that I show to be equivalent to unbiased estimation (see e.g. Mullainathan and Spiess, 2017). The proposed mechanism therefore allows researchers to leverage machine learning in estimating average treatment effects in experimental data. Wager et al. (2016) propose a similar estimator based on separate prediction problems in the treatment and control groups, and show that its asymptotic variance only depends on the prediction risk of the regression adjustments. Bloniarz et al. (2016) use the LASSO to select among control variables in experiments. Athey and Imbens (2016) use regression trees to estimate heterogeneous treatment effects. Chernozhukov et al. (2017a) estimate treatment effects from high-dimensional observational data. I contribute a finite-sample principal-agent framework for integrating machine learning, which is mostly agnostic about specific algorithms or asymptotic approximations.

My analysis is limited in three ways. First, I assume randomization, and thus that identification is resolved by design. My findings extend to known propensity scores, stratified and conditional randomization, and corresponding identification from quasi-experiments.[9] Second, I focus on the analysis of a single experiment, and neither on repeated interactions between designer and investigator, nor on the publication policies that may shape investigators' preferences. Third, I characterize optimal estimators in terms of prediction tasks, but I do not discuss in depth the solution to these prediction problems. A large and active literature that straddles econometrics, statistics, and machine learning provides guidance and tools to provide efficient prediction functions.

The remaining article is structured as follows. Section 1 introduces the main ideas behind my theoretical results in a stylized example. In Section 2, I formally lay out

[9]When treatment is not random, endogeneity creates auxiliary prediction tasks in the propensity score that interact with fitting regression adjustments (Robins and Rotnitzky, 1995; Chernozhukov et al., 2017b). Finite-sample unbiased estimation may then be infeasible absent strong parametric assumptions, and inference may be invalid when these additional prediction tasks are ignored (Belloni et al., 2014).

the specific estimation setting and my mechanism-design approach. I preview my main theoretical results in Section 3. In Section 4, I solve for optimal restrictions on the investigator's estimation. Section 5 characterizes unbiased estimators and solves for the investigator's second-best choice. For the case that full ex-ante commitment is infeasible or impractical, Section 6 considers unbiased estimators that permit ex-post researcher input. In the Conclusion, I discuss extensions. In the Appendix, I collect the proofs of my main results and discuss asymptotic inference. Additional proofs and supplementary results can be found in the Supplementary Appendix.

1 A Simple Example

I consider the estimation of a sample-average treatment effect. But the main features of my analysis are already apparent when we focus on a single unit within that sample. As an example, I discuss the estimation of the effect of random assignment to a job-training program on the earnings of one specific worker.[10]

1.1 Estimating the Unit-Level Causal Effect

The causal effect on unit i is $\tau_i = y_i(1) - y_i(0)$, where $y_i(1), y_i(0)$ are the potential outcomes when assigned to treatment or control, respectively. For assignment to a job-training program, $y_i(1) = \$1,190$ could be the earnings of worker i when he is offered the training program, and $y_i(0) = \$1,080$ the earnings of the *same* worker without access to this training, so $\tau_i = \$110$. We do not observe both potential outcomes for one unit simultaneously, but observe only the treatment status d_i and the realized outcome

$$y_i = \begin{cases} y_i(1), & d_i = 1, \\ y_i(0), & d_i = 0. \end{cases}$$

But since treatment is assigned randomly (with probability $p = \mathrm{P}(d_i = 1)$), we can still obtain an unbiased estimate of the unit treatment effect.[11] Indeed, I will

[10] Throughout, I focus on intent-to-treat effects, so I do not consider take-up or the use of random assignment as an instrument.

[11] Here, I assume that we know that treatment has been assigned with known probability $p = \mathrm{P}(d_i = 1)$. Throughout the remaining article, I also consider random assignment with a fixed number of treated units rather than a known ex-ante probability of treatment. The case of known number n_1 of treated units has structurally similar features, but is *not* the same as the case with known probability $p = \frac{n_1}{n}$. The reason for the difference is that knowledge of all *other* units'

argue below that $\frac{d_i - p}{p(1-p)} y$ is an unbiased estimator for τ_i. (Throughout, by "unbiased" I mean that, for fixed potential outcomes $y_i(1)$ and $y_i(0)$, the treatment-effect estimator averages out to $y_i(1) - y_i(0)$ over random draws of treatment d_i.)

In addition to the realized outcome y_i and treatment status d_i, I assume that we also have access to some pre-treatment characteristics x_i of unit i. Estimating the treatment effect $\tau_i = y_i(1) - y_i(0)$ for, say, a treated unit ($d_i = 1$) amounts to imputing the missing, counterfactual control outcome $y_i(0)$. When we have additional information about that unit, we can hope to use it together with the outcome, treatment, and characteristic data $z_{-i} = (y_j, d_j, x_j)_{j \neq i}$ of all other units to estimate $y_i(0)$, and thus τ_i. The investigator could, for example, run a linear regression of earnings on treatment, pre-assignment earnings, and some basic demographic characteristics to impute the counterfactual outcome $y_i(0)$. She could then estimate that worker's treatment effect by the difference between realized and imputed earnings.

If we do not put any restriction on estimation and investigator and social preferences agree, then the investigator's estimator will represent her expertise as well as the data. I model the investigator's expertise as a prior distribution π over potential outcomes $y_i(1), y_i(0)$ given characteristics x_i. (To be more precise, this prior will be over the joint distribution of the potential outcomes of all units given all their controls.) If the investigator aims to minimize the average mean-squared error $\mathrm{E}_\pi(\hat{\tau}_i - \tau_i)^2$, then for $d_i = 1$ she will estimate τ_i by

$$\hat{\tau}_i = \mathrm{E}_\pi[\tau_i | y_i, d_i, x_i, z_{-i}] = \underbrace{y_i(1)}_{\text{observed}} - \mathrm{E}_\pi[\overbrace{y_i(0)}^{\text{unobserved}} | y_i(1), x_i, z_{-i}].$$

This estimator represents the investigator's best guess of the treatment effect given her prior and all information in the data. In the training-program example, one specific prior could imply the use of Mincer polynomials in imputing the missing counterfactual outcome by its posterior expectation $\mathrm{E}_\pi[y_i(0) | y_i(1), x_i, z_{-i}]$.

1.2 Specification Searches and Optimal Restrictions on Estimation

If investigator and social preferences are misaligned, then the investigator's estimator may represent her incentives more than her expertise and the data. Even if the

treatment status is not informative about a given unit's treatment status for known p, but perfectly determines the left-out unit's treatment status for known n_1. Instead of leave-one-out regression adjustments, for fixed n_1 I therefore show in Section 5 that leave-two-out regression adjustments fully characterize unbiased treatment-effect estimators.

investigator commits to an estimator ex-ante, she could still choose one that is biased towards her preference rather than her prior. As the designer, we therefore should not only require that the investigator commits to an estimator before she has seen all of the data, but also restrict the estimators the investigator can choose from.

We face a tradeoff between flexibility and robustness. Constraints that are too permissive may lead to publication bias. One extreme solution would restrict the investigator to simple specifications that do not use control covariates, or use them only in simple linear regressions. Conventional pre-analysis plans often take this form. But restricting the investigator to a few estimators may forfeit experiment-specific knowledge about the relationship of control variables to outcomes in the prior, which I assume encodes the private information of the investigator.

I show that unbiasedness is a restriction on estimation that resolves this tradeoff. The first-best optimal estimator usually has bias. Indeed, the posterior expectation of the treatment effect τ_i is usually biased towards the investigator's prior expectation $\mathrm{E}_\pi \tau_i$. But when we leave the decision over bias to the investigator, then the investigator may shrink her estimator to her preferred estimate instead of her prior.

Once we restrict the investigator to unbiased estimators of τ_i, even an investigator who wants to minimize mean-squared error relative to some fixed target $\tilde{\tau}_i$ (rather than the true treatment effect) will minimize average mean-squared error relative to the true treatment effect among unbiased estimators, since the investigator's average risk (or cost in the nomenclature of mechanism design) is then

$$\mathrm{E}_\pi(\hat{\tau}_i - \tilde{\tau}_i)^2 = \underbrace{\mathrm{E}_\pi(\hat{\tau}_i - \tau_i)^2}_{\text{social preference}} + \overbrace{\mathrm{E}_\pi(\tau_i - \tilde{\tau}_i)^2}^{\text{unaffected by investigator choice}}.$$

My first main result is that fixing the bias (e.g. to zero) represents an optimal restriction in a minimax sense (Theorem 1) over a set of investigator preferences that generalize this risk function (Assumption 5). That is, the investigator's average mean-squared error is minimal for an investigator that minimizes mean-squared error relative to some worst-case target, given some (hyper-)prior over the investigator's private information.

1.3 Optimal Unbiased Estimation

Now that investigator and social preferences are aligned, how can the investigator choose an unbiased estimator with low variance? For the unit-level treatment effect τ_i, a simple unbiased estimator is available. Indeed, the estimator

$$\hat{\tau}_i = \frac{d_i - p}{p(1-p)} y_i = \begin{cases} +\frac{1}{p} y_i & d_i = 1, \\ -\frac{1}{1-p} y_i & d_i = 0, \end{cases}$$

is unbiased because $\mathrm{E}[\hat{\tau}_i] = p\frac{1}{p} y_i(1) - (1-p)\frac{1}{1-p} y_i(0) = \tau_i$. But this estimator can have very high variance. Assume that job training is assigned with probability $p = .5$, and that the potential earnings are $y_i(1) = \$1,190$ and $y_i(0) = \$1,080$. Then

$$\hat{\tau}_i = \begin{cases} +\$2,380 & d_i = 1, \\ -\$2,160 & d_i = 0, \end{cases}$$

is an unbiased, but extremely variable estimator of the treatment effect $\tau_i = \$110$. Indeed, the variance of $\hat{\tau}_i$ under treatment assignment is

$$\mathrm{Var}(\hat{\tau}_i) = p(1-p)(\hat{\tau}_i(d_i = 1) - \hat{\tau}_i(d_i = 0))^2,$$

so in the example the standard error amounts to $\sqrt{\mathrm{Var}(\hat{\tau}_i)} = \$2,270$.

We can modify this estimator by regression adjustments $\hat{y}_i$ to obtain

$$\hat{\tau}_i = \frac{d_i - p}{p(1-p)} (y_i - \hat{y}_i). \tag{1}$$

As long as $\hat{y}_i$ only uses information from x_i and $z_{-i} = (y_j, d_j, x_j)_{j \neq i}$, and not the outcome y_i or treatment effect d_i, $\hat{\tau}_i$ will still be unbiased. *My second main result* shows that all unbiased estimators of the treatment effect can be written in this way (Lemma 1). Concretely, any unbiased estimator of the sample-average treatment effect is the average over estimators $\hat{\tau}_i$ for all i that each include an adjustment that uses data only from all other units. All unbiased estimators are thus equivalent to a repeated sample-splitting procedure. Conversely, if $\hat{y}_i$ is fitted, for example, by a regression of y on x that violates the sample-splitting construction by also including y_i, then overfitting of $\hat{y}_i$ to y_i would bias the treatment-effect estimate towards zero.

Which regression adjustment minimizes variance? Optimally the investigator

would set $\hat{y}_i$ to $(1-p)y_i(1)+py_i(0)$, since this leads to $\hat{\tau}_i = \tau_i$. But without using $y_i(1)$ or $y_i(0)$, the investigator's best choice is the posterior expectation

$$\hat{y}_i = \mathrm{E}_\pi[(1-p)y_i(1)+py_i(0)|x_i, z_{-i}].$$

In the example, if the investigator's best guess of the expected potential earnings, $\frac{y_i(1)+y_i(0)}{2}$, based on her prior and data on all other units is $\hat{y}_i = \$1,100$, then

$$\hat{\tau}_i = \begin{cases} +2(\$1,190 - \$1,100) = \$180 & d_i = 1, \\ -2(\$1,080 - \$1,100) = \$40 & d_i = 0 \end{cases}$$

is still unbiased for $\tau_i = \$110$, but has much lower variance (the standard error is now $\sqrt{\mathrm{Var}(\hat{\tau}_i)} = \$70$). *My third main result* shows that the investigator's solution for the regression adjustments in general takes this form (Theorem 2), and as a corollary that all admissible (non-dominated) unbiased estimators can be achieved by exactly these regression adjustments (Theorem 3).

1.4 Machine Learning

By construction, the estimator in (1) of the unit-level treatment effect τ_i is unbiased whatever the regression adjustment is. In particular, the sample-splitting construction ensures that prior information only affects variance. Even a misspecified or dogmatic prior does not systematically bias what we learn about τ_i. This robust construction offers an opportunity to leverage tools that produce good predictions of potential outcomes even when they come with little guarantees that would otherwise ensure unbiasedness.

The optimal regression adjustments $\hat{y}_i = \mathrm{E}_\pi[(1-p)y_i(1)+py_i(0)|x_i, z_{-i}]$ solve an out-of-sample prediction problem. Take the special case $p = .5$.[12] Then $\hat{f}_i(x_i) = \mathrm{E}_\pi[.5y_i(1)+.5y_i(0)|x_i, z_{-i}]$ minimizes average prediction risk for the loss $(\hat{f}_i(x_i)-y_i)^2$ where $\hat{f}_i$ uses outcome and treatment data from all other units only. This is a regression problem where the quality of fit is measured at a new sample point, and not inside the training sample. Supervised machine-learning algorithms are built to solve exactly such out-of-sample prediction problems. For example, shrinkage methods like ridge regression of the LASSO can have better out-of-sample prediction

[12] When treatment is not balanced, $p \neq .5$, additional weights in the prediction loss express that adjustments for the smaller group effectively get weighted up in (1). For details, see (2) in Section 5.

performance than a linear least-squares regression that optimizes the in-sample fit.

I also obtain an intuitive formula for calculating standard errors. The variance of $\hat{\tau}_i$ is the expected loss in predicting the weighted potential outcome sum $(1-p)y_i(1)+py_i(0)$ by the adjustment $\hat{y}_i$, which can be estimated from the realized outcome y_i that has been excluded from the construction of $\hat{y}_i$. When units are sampled randomly, I show that, under mild conditions on the construction of regression adjustments, standard errors can be calculated from estimated prediction loss.

1.5 Unbiased Estimation without Pre-Specification

Regression adjustments incorporate flexibly the investigator's expertise as well as the data, but to ensure unbiasedness, the investigator must commit to their construction in advance. Indeed, once the investigator has seen the full sample data, she cannot credibly claim that some adjustment uses data only from other units. Practically, the investigator could pre-specify a machine-learning algorithm that learns regression adjustments from the data. But that may be impractical when the construction of adjustments requires input by the researcher.

However, complete pre-specification is not necessary to ensure unbiasedness. Instead the investigator could commit to splitting and distributing the sample. Assume there is a researcher in the investigator's research team that has not yet seen the data. To obtain a regression adjustment for unit i, the investigator could give that researcher access to data only from all other units. That researcher then takes the subsample, solves a prediction problem to obtain a good adjustment $\hat{y}_i$, and returns that regression adjustment to the investigator, who estimates the treatment effect according to (1). In that case, that researcher's choice will not introduce bias even if the researcher does not commit to the construction of the regression adjustments in advance.

Of course, estimating the average treatment effect on all n sample units in this way would require a team of n researchers. But my *fourth main result* characterizes all unbiased estimators that remain feasible without detailed pre-specification and when only K researchers are available (Corollary 1). Even ex-post analysis by a single researcher improves over simple pre-analysis plans without the need for detailed pre-specification. I also show that delegating estimation to two researchers approximates optimal estimation in that it ensures asymptotic efficiency under mild conditions.

2 Setup

Having given a simple example, I now lay out formally how I approach causal inference as a mechanism-design problem. A designer delegates the estimation of an average treatment effect in a randomized experiment to an investigator. The investigator receives a private signal about the distribution of potential outcomes, but has unknown preferences that can be biased. The designer does not analyze the dataset herself, but instead sets constraints on the investigator's estimator.

In this section, I first define the data-generating process and target parameter before introducing the investigator's and designer's problems. To simplify the further analysis, I then argue that we can restrict the analysis to direct restrictions by the designer on the space of estimators the investigator commits to.

2.1 Target Parameter

Following Neyman (1923), I am interested in the average treatment effect

$$\tau_\theta = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(y_i(1) - y_i(0))}_{=\tau_i} \qquad\qquad \theta = (y_i(1), y_i(0))_{i=1}^n$$

in a given sample of n units. In the Rubin (1974, 1975, 1978) causal model interpretation, $y_i(d_i)$ is the potential outcome of unit i had they received treatment status $d_i \in \{0, 1\}$, and τ_i the respective causal effect.

The n units may be randomly sampled from a population distribution,

$$(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P},$$

with pre-treatment characteristics $x_i \in \mathcal{X}$. In this case, my analysis will extend to the estimation of the population-average treatment effect $\tau = \mathrm{E}[y_i(1) - y_i(0)]$ and the conditional average treatment effect (given characteristics $x \in \mathcal{X}^n$) $\frac{1}{n}\sum_{i=1}^n \mathrm{E}[y_i(1) - y_i(0)|x_i]$. My main analysis is conditional on $(y_i(1), y_i(0), x_i)_{i=1}^n$ and therefore focuses on the sample-average treatment effect τ_θ, but I will return to τ when I discuss inference.

2.2 Experimental Setup

I assume that treatment is assigned randomly to overcome the missing-data problem central to causal inference (Holland, 1986). For a unit with treatment status d_i, we only observe the realized outcome $y_i = y_i(d_i)$. But because I assume that the distribution of treatment assignment $d \in \{0,1\}^n$ does not vary with the potential outcome vectors $y(1), y(0) \in \mathbb{R}^n$ (Cochran, 1972), we can estimate the treatment effect without bias. The stable-unit treatment effect assumption (Rubin, 1978) of no interference between units is implicit.

Assumption 1 (Random Treatment)**.** *Given potential outcomes $\theta = (y_i(1), y_i(0))_{i=1}^n$, the data $z = (y_i, d_i)_{i=1}^n$ is distributed according to P_θ as follows. d is generated from a known distribution over $\{0,1\}^n$ that does not depend on $(y(1), y(0))$ and is one of:*

1. *Each unit is independently assigned to treatment with known probability $p = \mathrm{P}(d_i = 1)$ (where $0 < p < 1$).*
2. *d is drawn uniformly at random from all assignments with known number $n_1 = \sum_{i=1}^n d_i$ of treated units (where $0 < n_1 < n$).*

Given d, $y_i = y_i(d_i)$ for all $i \in \{1, \ldots, n\}$.

In this notation, I do not explicitly include the covariates $x_1, \ldots, x_n$ in the data z, since I condition on the controls and therefore treat $(x_i)_{i=1}^n$ as a constant and not as a random variable. While neither of the distributions of d depends on the controls, my results will extend to distributions that are known functions of x_i if they ensure identification of τ_θ. These include stratified or conditional random sampling, and sampling according to known propensity scores.

2.3 Covariate Adjustments

How can we estimate the sample-average treatment effect τ_θ from data $(y_i, d_i, x_i)_{i=1}^n$? Since treatment is exogenous, the average difference

$$\hat{\tau}^*(z) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j) = \frac{1}{n_1} \sum_{d_i=1} y_i - \frac{1}{n_0} \sum_{d_i=0} y_i$$

between treatment and control outcomes is an unbiased estimator of τ_θ conditional on the number n_1 of treated units (provided $0 < n_1 < n$).

Of course this difference in averages $\hat{\tau}^*$ leaves information in the covariates $x_1, \ldots, x_n$ on the table and is likely inefficient. In econometric practice, τ_θ is therefore often estimated from a linear regression of the outcome on treatment and controls. But the researcher's choice of control strategy can bias published results. First, implicit model assumptions may bias estimates. Even simple linear regressions can be biased (Freedman, 2008), although this bias vanishes asymptotically if interactions are included (Lin, 2013). Second, if the investigator does not document that she picked among multiple covariate adjustments, an unsuspecting observer's inference may be biased towards stronger treatment effects and unjustified confidence (Lenz and Sahn, 2017).

2.4 Estimation Preferences

I explicitly consider the choice of the control specification in a mechanism-design framework. A designer and an investigator face a choice of an estimator

$$\hat{\tau} : \mathcal{Z} \to \mathbb{R}$$

that maps experimental data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ into an estimate $\hat{\tau}(z)$ of the sample-average treatment effect τ_θ. Since my analysis is conditional on the control covariates, this estimator encodes in particular how the estimate of the treatment effect is adjusted for the realizations $x_1, \ldots, x_n$ of the control variables.

Designer and investigator preferences are expressed by risk functions $r^D, r^I : \Theta \times \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$ that encode the expected loss $r^D_\theta(\hat{\tau}), r^I_\theta(\hat{\tau})$ of an estimator $\hat{\tau} \in \mathbb{R}^{\mathcal{Z}}$ given the full matrix $\theta = (y(1), y(0)) \in \mathcal{Y}^{2n} = \Theta$ of $2n$ potential outcomes in the sample at hand. Both designer and investigator aim to minimize their respective risk given the potential outcomes θ. Throughout this article, I specifically assume that the designer's risk function expresses a social desire to obtain precise estimates of the true treatment effect τ_θ.

Assumption 2 (Social risk function)**.** *The designer's risk for an estimator $\hat{\tau} : \mathcal{Z} \to \mathbb{R}$ is the estimator's mean-squared error*

$$r^D_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tau_\theta)^2],$$

where the expectation averages over random treatment assignment given potential outcomes $\theta \in \Theta$.

Notably, I do not assume that the designer has an inherent preference for unbiased estimators.[13] While my characterization results will depend on this specific form of the social risk function, the general mechanism-design approach extends to alternative risk (or equivalently utility) functions.

The investigator's risk function can differ from the designer's risk function. For example, I will later consider risk functions that include $r^I_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau})^2]$, which expresses a desire to obtain a certain estimate $\tilde{\tau}$ irrespective of the true treatment effect τ_θ. The designer knows only that $r^I \in \mathcal{R}$ for some set of risk functions.

2.5 Prior Information

Since generally no single estimator $\hat{\tau}$ minimizes risk for all potential outcomes $\theta \in \Theta$ and θ is not known, a good estimator has to trade off risk performance across different draws of potential outcomes. Following Wald (1950), I assume that a prior distribution π over potential outcomes governs this tradeoff.[14]

The investigator receives the prior distribution π over potential outcomes θ as a private signal before the data z is realized. This private information models researcher expertise. For example, the investigator may have run previous studies or a pilot and synthesize relevant results in the literature. The investigator therefore has a sense which variables are important and which regression specifications are more likely to work well.

The uninformed designer does not observe the prior π, but only has a diffuse (hyper-)prior η for π. The designer therefore designs a mechanism that elicits the investigator's prior information. Optimally, the designer would want to obtain an estimator that minimizes average mean-squared error given the investigator's private prior, but since the investigator's preferences may differ from the designer's, the latter cannot generally achieve a first-best estimator.

[13] Still, the minimization of squared-error loss is associated with unbiasedness, as e.g. in Lehmann and Romano (2006, Example 1.5.6).

[14] One alternative approach to finding a good estimator would involve putting restrictions on the distribution of potential outcomes and discussing efficient estimators under some large-sample approximation. But since researchers may reasonably disagree about these choices, this would itself add an additional degree of freedom to estimation. I instead consider estimation in an exact finite-sample decision-theoretic framework that does not restrict the distribution of potential outcomes.

2.6 Mechanism Structure and Timeline

I assume that the designer has the authority to set rules in the form of a mechanism without transfers. The designer cannot verify the investigator's risk type or private prior information. The investigator follows whatever mapping from investigator decisions to final estimator the designer sets, and the designer follows through on the mapping she commits to. Similar to Frankel's (2014) delegation setup, the game between designer and investigator plays out in the following steps:

1. The designer chooses a mechanism that consists of a message space M and a mapping from messages m into estimators $\hat{\tau}_m : \mathcal{Z} \to \mathbb{R}$.

2. The investigator observes the prior distribution π and sends a message $m(r^I, \pi)$.

3. The potential outcomes θ are realized, the data z drawn according to the experiment, and the estimate $\hat{\tau}_{m(r^I,\pi)}(z)$ formed.

In econometric terms, I think of the investigator's message as a modelling decision. The designer then restricts the space of models the investigator can choose from.

For simplicity, I assume that the investigator's message given her risk type and private information and the mapping of her message to the final estimator are deterministic, but the setup extends to stochastic actions as in Frankel (2014). By the revelation principle, the specific form of the mechanism is not a substantial restriction, since it includes direct mechanisms in which the investigator reveals her risk type and her private information (as e.g. in Holmström, 1984).

Since the investigator controls the estimator with her choice of message, we can assume without loss of generality that the message space is a set of estimators (and the mapping from message to estimator the identity). Indeed, take any estimator that is an outcome for some message. Since neither risk type nor prior are verifiable, the investigator can always choose that message to obtain said estimator.

Hence, the designer directly restricts estimators to some set $\mathcal{C}^D$. Subject to the constraint, the investigator specifies an estimator $\hat{\tau}^I \in \mathcal{C}^D$ before data becomes available. Once the data $z \in \mathcal{Z}$ is realized, the investigator reports the estimate $\hat{\tau}^I(z)$ (Figure 1). Since my econometric analysis is conditional on the control variables $x_1, \ldots, x_n$, this baseline information can be available to the investigator and inform her choice of estimator.

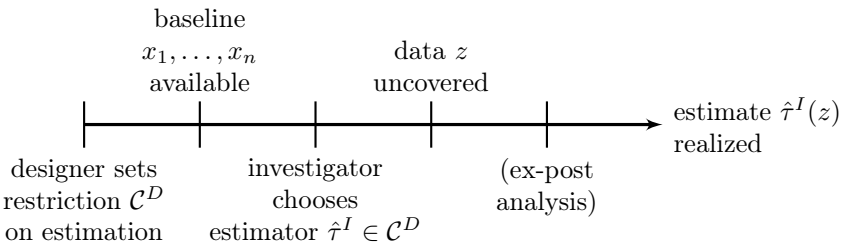

Figure 1: Estimation timeline

Optimal estimation in this framework will require some degree of commitment by the investigator before the data is available. Otherwise, any restriction on estimation would be cheap talk, since the investigator could choose an estimator ex post that justifies their preferred estimate at the realized data. But I will show that optimal commitment is less constraining than restricting the investigator to pre-analysis plans with simple specifications that are chosen ex ante. First, the investigator's estimator can still contain (automated) specification searches. Second, in Section 6, I show that it is not generally necessary to specify the full estimator ex ante, and that additional exploratory analysis after the data has become available can improve estimation.

2.7 Investigator and Designer Choices

Having set up the actions available to the investigator and designer, I now describe their preferences. The investigator chooses an estimator to minimize average risk subject to her prior.

Assumption 3 (Investigator's choice)**.** *Given the prior distribution π over potential outcomes $\theta \in \Theta$, the investigator minimizes average risk subject to the constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ set by the designer,*

$$\hat{\tau}^I = \hat{\tau}^I(\mathcal{C}^D, \pi) \in \arg\min_{\hat{\tau} \in \mathcal{C}^D} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})].$$

The designer does not know the risk function of the investigator, but only assumes that it falls within some set $\mathcal{R}$ of risk functions. Adapting the maxmin criterion from the mechanism-design literature (e.g. Hurwicz and Shapiro, 1978; Frankel, 2014; Carroll, 2015), I assume that the designer chooses a constraint that minimizes average risk at a worst-case investigator type within that set.

Definition 1 (Designer's minimax delegation problem)**.** *Given some set $\mathcal{R}$ of investigator risk functions, the designer picks a constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ to minimize average mean-squared error,*

$$\mathcal{C}^D = \mathcal{C}^D(\mathcal{R}, \eta) \in \min_{\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}} \sup_{r^I \in \mathcal{R}} \mathrm{E}_\eta[r^D_\theta(\hat{\tau}^I)],$$

where I assume that the investigator breaks ties in the designer's favor.

The minimax criterion can be seen as a game between designer and nature. For every choice of restriction that the designer picks, nature responds with an investigator who produces maximal average mean-squared error. In this game, the designer picks a constraint that ensures that the average risk at a worst-case outcome is minimal.

Without constraints, the investigator's estimator may be a poor fit from the designer's perspective. But if the constraints are too restrictive, for example if we reduce the allowed set of estimators to the difference in averages $\hat{\tau}^*$, we will use the investigator's expertise inefficiently. I therefore solve for constraints $\mathcal{C}^D$ that resolve this tradeoff between flexibility and robustness optimally.

2.8 Support Restriction

Throughout this article, I assume that the support of (potential) outcomes is finite, for three reasons. First, I adapt results from the mechanism-design literature that involve finite sums. Second, I use and provide complete-class theorems that fully characterize admissible (non-dominated) estimators provided their support is finite. Third, I derive intuitive combinatorial proofs for my characterization results.

Assumption 4 (Finite support)**.** *The support $\mathcal{Y}$ of potential outcomes $y_i(1), y_i(0)$ is finite.*

Since the number of support points is otherwise unrestricted, the finite-support assumption allows for flexible approximations to arbitrary distributions.

3 Overview of Main Results

In this section, I preview my main theoretical results. Under reasonable restrictions on investigator preferences, I show that fixing the bias is a minimax optimal constraint on estimation. I then present a representation of unbiased treatment-effect

estimators, characterize the investigator's optimal choice from this restricted class, and extend the analysis to estimators with limited pre-specification.

I assume that investigator risk functions express mean-squared error relative to some target which may not be the true treatment effect.

Assumption 5 (Investigator risk restriction)**.** *The investigator has a risk function from the set*

$$\mathcal{R}^* = \{r^I; r^I_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \to \mathbb{R}\}.$$

The target function $\tilde{\tau}_\theta$ is unrestricted in this definition. For example, the investigator may want to achieve a constant target no matter what the true potential outcomes are ($\tilde{\tau}_\theta = \text{const.}$). Or the investigator may prefer to obtain estimates above the true treatment effect ($\tilde{\tau}_\theta = \tau_\theta + \varepsilon$).

In any of these cases, restricting investigators to unbiased estimators ($\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$) ensures that they choose among these estimators as if they had the designer's preference, i.e. they minimize average variance. Once I have established tools for asymptotically valid inference, it will also follow that unbiasedness aligns the choices of investigators who want to obtain a small standard error or a low p-value.

While fixing the bias aligns preferences, this restriction may be too strong. However, I establish that it is minimax optimal for an appropriate choice of biases.

Theorem 1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r^D_\eta \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})] \right) \right].$$

This result implies that the designer should not leave the choice of bias to the investigator. If the designer has an informative prior, she may set biases to reflect that information. But with little information on the designer's side, a natural choice of biases is zero.

With a zero-bias restriction, the investigator chooses among all unbiased estimators to minimize variance. The next result characterizes all unbiased estimators, and therefore the choice set of the investigator.

Lemma 1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

1. *For a known treatment probability p, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}'d_{-ij} = n_1 - 1$.

All unbiased estimators are hence sample-splitting estimators that leave one or two units out, respectively, when calculating their regression adjustments. But when is an estimator not just unbiased, but also precise? The investigator would optimally want to set regression adjustments to the oracle solutions

$$\begin{aligned} \bar{y}_i &= (1-p)y_i(1) + py_i(0), \\ \Delta\bar{y}_{ij} &= \left(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\right) - \left(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\right), \end{aligned}$$

respectively, but since the potential outcomes are unknown, these adjustments are infeasible. Instead, I show that the investigator chooses leave-one-out or leave-two-out expectations of these adjustments.

Theorem 2 (Solution of the investigator)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i | z_{-i}]).$$

2. For a fixed number n_1 of treated units,

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta \bar{y}_{ij} | z_{-ij}]).$$

Hence, all optimal unbiased estimators take as regression adjustments conditional expectations of potential outcomes. These conditional expectations can be obtained as solutions to a prediction problem. Independently of the mechanism-design setup, the set of investigator solutions across different priors completely characterize the class of admissible unbiased estimators of the sample-average treatment effect.

Theorem 3 (Complete-class theorem for unbiased estimators)**.** *For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ with full support such that $\hat{\tau}$ equals the limit of the respective estimators in Theorem 2. Conversely, for any converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ that put positive weight on every state $\theta \in \Theta$, the limit of the estimators is admissible among unbiased estimators.*

Now that I have characterized the optimal solution of the designer and the investigator, I return to the question of commitment. The representation of unbiased estimators in Lemma 1 requires that the construction of regression adjustments does not involve the adjusted unit. In Theorem 2, the investigator would therefore have to commit to their construction before she has access to the full sample. This pre-specification leaves room for automated specification searches in constructing the adjustments. But fully pre-specifying all specification searches may be impractical.

I also characterize estimators that ensure unbiasedness not by the investigator fully pre-specifying adjustments, but by a commitment to a sample-splitting scheme. I consider estimation contracts that have the investigator delegate estimation tasks on subsamples to K researchers who do not share information about the data they receive.

Definition 2 (K-distribution contract)**.** *A K-distribution contract $\hat{\tau}^\Phi$ distributes data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ to K researchers. Researcher k receives data $g_k(z) \in A_k$ and returns the intermediate output $\hat{\phi}_k(g_k(z)) \in B_k$. The estimate is*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \Phi((\hat{\phi}_k(g_k(z))_{k=1}^K; z).$$

The investigator chooses the functions g_k (from data in $\mathcal{Z}$ to researcher input in A_k)

and Φ *(from the researcher outputs in* $\bigtimes_{k=1}^{K} B_k$ *and data in* $\mathcal{Z}$ *to estimates in* $\mathbb{R}$*) before accessing the data.*

As one special case of my general representation result of unbiased K-distribution contracts, I characterize unbiased estimators that divide the sample into K folds and then give each researcher access to all but one of these folds. In that case, I deduce from the representation of unbiased estimators in Lemma 1 that the estimator is unbiased if and only if each researcher only controls the regression adjustments for the respective left-out fold.

Corollary 1 (Characterization of unbiased K-fold distribution contracts)**.** *For* K *disjoint folds* $\mathcal{I}_k \subseteq \{1, \dots, n\}$ *with projections* $g_k : (y, d) = z \mapsto z_{-\mathcal{I}_k} = (y_i, d_i)_{i \neq \mathcal{I}_k}$*, a* K*-distribution contract* $\hat{\tau}^{\Phi}$ *is unbiased if and only if:*

1. *For a known treatment probability* p*,there exist a fixed unbiased estimator* $\hat{\tau}_0(z)$ *and regression adjustment mappings* $(\Phi_k)_{k=1}^K$ *such that*

$$\hat{\tau}^{\Phi}((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z_{-i})$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$*.*

2. *For a fixed number* n_1 *of treated units, there exist a fixed unbiased estimator* $\hat{\tau}_0(z)$ *and regression adjustment mappings* $(\Phi_k)_{k=1}^K$ *such that*

$$\hat{\tau}^{\Phi}((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^{K} \sum_{\{i<j\} \subseteq \mathcal{I}_k} (d_i - d_j) \phi_{ij}^k(z_{-ij}),$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$*.*

These sample-distribution contracts achieve unbiasedness without detailed commitments by the researchers. For $K = 1$, I show that giving one researcher (with risk function in the set $\mathcal{R}^*$) access to part of the sample for exploratory ex-post analysis can improve over simple pre-analysis plans. For $K = 2$, I show that a flexible pre-analysis plan that specifies distribution to two researchers asymptotically achieves semi-parametric efficiency when the units are sampled iid under conditions on the population distribution.

4 Designer's Solution

Having set up the estimation of a sample-average treatment effect as a mechanism-design problem, I justify a restriction to unbiasedness by solving the designer's delegation problem. Subject to unbiasedness, the investigator pre-specifies an estimator according to the designer's preferences. I prove minimax optimality of fixed-bias restrictions, echoing a result from mechanism design on optimal delegation.

4.1 The Role of Bias

When there is no misalignment of preferences, then the resulting first-best estimator that minimizes average mean-squared error will generally have bias. To understand how being flexible on bias can improve estimation, note that both bias and variance contribute to the risk

$$r_\theta^D(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau} - \tau)^2] = \underbrace{(\mathrm{E}_\theta[\hat{\tau}] - \tau)^2}_{\text{bias}} + \underbrace{\mathrm{Var}_\theta(\hat{\tau})}_{\text{variance}}$$

the designer aims to minimize. We can often improve an unbiased estimator by moving along this bias-variance tradeoff. Indeed, consider the first-best solution $\hat{\tau}_\pi = \arg\min_{\hat{\tau}} \mathrm{E}_\pi[r_\theta^D(\hat{\tau})]$ of the designer. The estimate $\hat{\tau}_\pi(z) = \mathrm{E}_\pi[\tau_\theta|z]$ comprises the posterior expectations $\mathrm{E}_\pi[y_i(1) - y_i(0)|z]$, which are usually biased towards the prior expectation of unit treatment effects when the prior is informative along this dimension.

But if the designer leaves the decision over bias to the investigator, then an investigator who has biased preferences will be inclined to bias the estimator in the direction of her preferences, not of her prior. Consider an investigator with risk

$$r_\theta^I(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - (\tau_\theta + \varepsilon))^2] \qquad (\varepsilon > 0)$$

who would like to show that the treatment effect is higher than it is. The investigator's unconstrained solution is now shifted upward by ε, which is added to the bias term. While reducing the variance relative an unbiased estimator, the designer's risk may also be increased through additional bias.

For choices among estimators with fixed bias, however, the investigator's and designer's preferences in this example are perfectly aligned. With bias fixed at zero, say, mean-squared error is variance, $r_\theta^D(\hat{\tau}) = \mathrm{Var}_\theta(\hat{\tau})$. The ε-biased investigator's

risk is $r_\theta^I(\hat{\tau}) = \varepsilon^2 + \text{Var}_\theta(\hat{\tau})$. While risks are not the same, they are shifted by a constant. There is no distortion in choices between estimators with fixed bias for this investigator loss function.

4.2 Unbiased Estimation as Second-Best

Having motivated in an example that fixing the bias can align investigator choices, I extend alignment to a minimax result. If the investigator has constant bias, I have shown that among estimators with fixed bias she will still commit to a variance-minimizing estimator. To show that this example extends to an optimal solution, I have to establish that the unbiasedness restriction is neither too permissive nor too restrictive.

The unbiasedness restriction

$$\mathcal{C}^* = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \text{E}_\theta[\hat{\tau}] = \tau_\theta \forall \theta \in \Theta\}$$

is not too permissive provided that investigators all choose as if they minimized mean-squared error relative to *some* target, albeit not necessarily relative to the true treatment effect.

Assumption 5 (Investigator risk restriction)**.** *The investigator has a risk function from the set*

$$\mathcal{R}^* = \{r^I; r_\theta^I(\hat{\tau}) = \text{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \to \mathbb{R}\}.$$

The target $\tilde{\tau}_\theta$ can vary arbitrarily with the potential outcomes. In particular, permissible risk functions include constant biases relative to the truth ($\tilde{\tau} = \tau + \varepsilon$) or fixed estimation targets ($\tilde{\tau} = \text{const.}$). $\mathcal{R}^*$ also includes the designer's risk function r^D at $\tilde{\tau} = \tau$.

Lemma 4.1 (Unbiasedness aligns estimation)**.** *If the investigator has risk from $\mathcal{R}^*$ then the investigator will choose from the unbiased estimators $\mathcal{C}^*$ according to the designer's preferences.*

Once I have established asymptotically valid inference for unbiased estimators in Appendix E, I will also show in Remark E.3 that the unbiasedness restriction aligns the choices of investigators who want to obtain small standard errors or tight confidence intervals. For a local-to-null alternative, by Remark E.4 unbiasedness

also insures asymptotic alignment in large samples when the investigator wants to obtain a low p-value (that is, wants to maximize the power of a test against some null hypothesis $\tau_\theta = \tau_0$).

Note, however, that there are many risk (or equivalently utility) functions for which unbiasedness does not provide alignment. In particular, unbiasedness may be a poor alignment device for non-convex loss functions. Take an investigator who wants to produce an estimate that does *not* reject some null hypothesis, for example when running a balance or robustness check. In that case, if some valid way of calculating standard errors is available, the investigator would want to obtain high variance even among unbiased estimators in order to weaken the evidence against her preferred null hypothesis.

For the class $\mathcal{R}^*$ of investigator risk functions, fixing the bias is not too restrictive because it is minimax optimal over investigator preferences. While Lemma 4.1 establishes that choices from unbiased estimators will be the same for any $r^I \in \mathcal{R}^*$, there could be a larger set of estimators that provide alignment, or full alignment of preferences could be too costly.

Theorem 1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r^D_\eta \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi [r^I_\theta(\hat{\tau})] \right) \right].$$

This minimax result shows that the gains from variance reduction of being flexible on bias are fully undone by the cost of misalignment for a worst-case risk function, for any relaxation of the fixed-bias restriction. Once we allow the bias to track the prior, it could as well reflect the preference of a worst-case investigator. The designer therefore chooses fixed biases that reflect her hyperprior.

Setting $\beta_\theta \equiv 0$ (and thus $\mathcal{C}^\eta = \mathcal{C}^*$) represents a hyperprior that is uninformative about the location of treatment effects. In principle, if the designer has a hyperprior η that is quite informative, she could introduce biases towards expected treatment effects under that hyperprior. Crucially, however, these biases would be fixed ex-

ante and not chosen by the investigator. In Supplementary Appendix F, I construct hyperpriors that deliver zero bias. There I also argue how an approximately uninformative hyperprior delivers approximately unbiased estimation in general as the support grows.

4.3 Connection to Aligned Delegation

My econometric finding that unbiased estimation is minimax optimal (Theorem 1) builds upon a mechanism-design result by Frankel (2014). There, a principal delegates decisions to an agent who observes states. Frankel (2014) characterizes optimal delegation mechanisms without transfers. In a class of maxmin optimal, simple mechanisms, the agent behaves according to the principal's preferences.

In a leading example from Frankel (2014), a school principal delegates the grading of a group of students to a teacher. The teacher may prefer to give more skewed or better grades than the principal, who does not observe the students' performance. However, the principal can exploit that the teacher's biased preferences are consistent across students. If the teacher and the principal agree on the ranking of students, fixing the distribution of grades obtains a second-best grade assignment. If the teacher has a constant bias, fixing the average grade already achieves agreement between principal and teacher. In both cases, the teacher chooses from the restricted grade assignments according to the principal's preferences.

What a fixed average is to grading in Frankel (2014), constant bias is to estimation in my setting. More precisely, I identify Frankel's (2014) school principal with my designer, the teacher with the investigator, and individual students with different draws of the data. In the school example, the performance of students is the private information of the teacher. For estimation, the prior distribution over potential outcomes is the private information of the investigator. Where the teacher chooses a grade for each student, the investigator commits to an estimator, that is, the investigator chooses an estimate for each (potential) draw of the data.

Frankel (2014) shows that fixing the average over grades is a maxmin (in utility terms) optimal mechanism for a class of biased squared-error preferences. Analogously, my fixed-bias restriction fixes weighted sums over estimates. But since fixing the bias requires setting many sums at once, and the designer's and investigator's preferences involve weights determined by the prior, additional work is required to establish the minimax optimality in Theorem 1. In Appendix A, I show how Frankel's (2014) result carries over to the designer's problem across all $\theta \in \Theta$, where

the investigator sets all $(2|\mathcal{Y}|)^n$ values of $\hat{\tau}(y,d)$ simultaneously.

4.4 Design of Experiment vs Design of Estimator

In Theorem 1, I have assumed that treatment is assigned randomly according to some fixed rule, but my results extend to the design of treatment assignment itself. The investigator may leverage prior knowledge about potential outcomes to adjust propensity scores (Kasy, 2016). For example, if the prior distribution of treated outcomes has larger variance than that of controls, the investigator may want to assign more units to treatment. Under the unbiasedness restriction, the investigator's preference over this additional decision remains aligned with the goal of the designer.

5 Investigator's Solution

The designer restricts the investigator to unbiased estimators. In solving the investigator's constrained optimization problem, I establish that optimal unbiased estimation is equivalent to a set of out-of-sample prediction tasks. I obtain a complete-class theorem that characterizes admissible unbiased estimators of the sample-average treatment effect.

Throughout this section, I assume that the investigator fully specifies her estimator before it is applied to outcome and treatment data $z = (y,d)$. Although the estimator is pre-specified, it can still include (automated) specification searches. The pre-specified estimator thus plays the role of a flexible pre-analysis plan. Since my results hold conditional on potential outcomes, the covariates $x_1,\ldots,x_n$ can be common knowledge before this pre-analysis plan is filed. In Section 6, I show how the results in this section extend when pre-specification is impractical. There, I provide a constructive characterization of pre-analysis plans that only commit to the way the sample is split and distributed.

5.1 Characterization of Unbiased Estimators

When is an estimator unbiased, conditional on potential outcomes? The designer requires that the investigator provides an unbiased estimator. In this section, I provide an intuitive representation of unbiased estimators that the investigator can achieve transparently by construction.

A class of estimators that ensures unbiasedness is obtained by sample splitting. For known treatment probability p, the Horvitz and Thompson (1952) estimator

$\hat{\tau}^{\text{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i$ is unbiased for any pair of potential outcome vectors because

$$\mathrm{E}_\theta\left[\frac{d_i - p}{p(1-p)} y_i\right] = y_i(1) - y_i(0).$$

If we replace outcomes y_i by adjusted outcomes $y_i - \phi_i(z_{-i})$ with regression adjustments that do not vary with (y_i, d_i), where z_{-i} denotes the data $(y_j, d_j)_{j \neq i}$ from all units other than i, then the resulting estimator is still unbiased.[15] (Recall that I condition on controls $x_1, \ldots, x_n$ throughout.) Since the adjustment $\phi_i(z_{-i})$ is the same whether unit i is treated or not and $\mathrm{E}_\theta\left[\frac{d_i - p}{p(1-p)} \middle| z_{-i}\right] = 0$, their addition averages out to zero, no matter the potential outcomes or realized treatment of the other units.[16]

I show that these sample-splitting estimators are also all estimators that are unbiased conditional on potential outcomes. If an estimator cannot be written as a Horvitz and Thompson (1952) estimator with leave-one-out regression adjustments, it must have bias for some matrix of potential outcomes. If instead we considered estimators that are unbiased given some distribution of potential outcomes (for example, we may want to model noise terms in potential outcomes that we do not want to condition on), then the result would trivially extend as long as we do not restrict this distribution. If an estimator cannot be written in this leave-one-out form, it must have bias for some distribution of potential outcomes.

A leave-one-out estimator can have bias conditional on the number of treated units. If the number n_1 of treated units is known, the leave-one-out adjustment $\phi_i(z_{-i})$ implicitly depends on $d_i = n_1 - \sum_{j \neq i} d_j$. For permutation randomization, I therefore start with the difference in averages $\hat{\tau}^* = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j)$ and establish that all unbiased estimators differ from $\hat{\tau}^*$ only by leave-two-out regression adjustments $\phi_{ij}(z_{-ij})$. In every sample split, these unbiased estimators leave out one treated and one untreated unit.[17]

Lemma 1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

[15] When adjustments are constructed from estimated potential outcomes, Wu and Gagnon-Bartsch (2017) call the resulting estimator for known p the "leave-one-out potential outcomes" (LOOP) estimator. This estimator is a special case of Aronow and Middleton's (2013) modification of the Horvitz and Thompson (1952) estimator.

[16] It would not be enough to exclude the treatment status d_i from the constriction of unit i's regression adjustment, and thus use y_i, since y_i can be correlated with d_i.

[17] Wager et al. (2016) consider leave-one-out estimators separately in the treatment and control groups, and use a leave-two-out construction to derive asymptotic unbiasedness.

1. *For a known treatment probability p, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}' d_{-ij} = n_1 - 1$.

The representations are restrictive, but not unique. In the minimal non-trivial case $n = 2$ and $|\mathcal{Y}| = 2$ for known treatment probability, the leave-one-out representation reduces the dimension of estimators $\hat{\tau} \in \mathbb{R}^{(\mathcal{Y} \times \{0,1\})^n}$ from 16 to 8. The unbiased estimators form a 7-dimensional affine linear subspace, and equivalent representations lie on lines in Euclidean space.

Notably, linear regression can not generally be represented in this way, as it is not generally unbiased in my setting (Freedman, 2008). In Appendix D, I provide a simple example of a biased OLS regression. Also, I make a connection between overfitting and bias, and show that bias can persist even under sampling from a population distribution and in large samples with high-dimensional controls.

We usually associate sample splitting with losses in efficiency in return for robustness. Since all unbiased estimators must split the sample, this logic applies here only through the robustness of the unbiasedness assumption to any distribution of potential outcomes. As long as we do not impose additional structure, all admissible (with respect to variance or equivalently mean-squared error) unbiased estimators must be among the sample-splitting estimators.

This result implies that the set of unbiased estimators the investigator chooses from is characterized by prohibitions. When we represent an estimator by a sum over adjusted outcomes, then there must be one such representation for which the investigator is not allowed to use the outcome and treatment assignment of a unit to construct its adjustment. For this prohibition to apply, in practice the investigator has to commit how the adjustment is constructed before she has access to the respective outcome and treatment status. I show below that this commitment leaves

room for automated specification searches, and discuss in Section 6 that human specification searches also remain feasible.

5.2 Solution to the Investigator's Problem

Given the unbiasedness restriction, what is the optimal solution of the investigator? The sample-splitting representation provides an objective criterion for unbiasedness. Since preferences are aligned, the investigator applies their subjective prior to minimize average risk among unbiased estimators. I characterize the resulting Bayes estimator by the solution to prediction problems.

The investigator solves a variance-minimization problem over the regression adjustments from Lemma 1. If the investigator knew the potential outcomes, a set of variance-minimizing regression adjustments would be given by the infeasible oracle solutions

$$\begin{aligned} \bar{y}_i &= (1-p)y_i(1) + py_i(0), \\ \Delta\bar{y}_{ij} &= \underbrace{\Big(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\Big)}_{=\bar{\bar{y}}_i} - \Big(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\Big) = \bar{\bar{y}}_i - \bar{\bar{y}}_j. \end{aligned}$$

I establish that the respective Bayesian leave-one-out and leave-two-out posterior expectations minimize average risk.[18] The resulting estimator is a constrained Bayes estimator in the sense of Wald (1950).

Theorem 2 (Solution of the investigator)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i|z_{-i}]).$$

2. *For a fixed number n_1 of treated units,*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta\bar{y}_{ij}|z_{-ij}]).$$

[18] In the case of known p this is similar to Wu and Gagnon-Bartsch's (2017) LOOP estimator, which estimates $y_i(1)$ and $y_i(0)$ separately from all other units and then averages these estimates with weights $1-p$ and p to obtain an adjustment that estimates $\bar{y}_i$.

The theorem is non-trivial because one adjustment appears in the estimate for multiple draws of the data. In particular, if two sample draws only differ in one unit, then the adjustments to that unit are the same. Key to the proof (which I develop in Appendix C) is solving a system of first-order conditions jointly for all potential draws of the data.

While the objective unbiasedness restriction dictates sample splitting and guarantees preference alignment, the prior picks one suitable estimator that trades off risk optimally between different unobserved states. If the prior assigns low probability to the realized set of potential outcomes, then the estimator is still unbiased, but may have high variance. In any case, the investigator wants to reveal her best guess given prior knowledge.

Sample splitting guards not just against misaligned preferences, but also against priors that are dogmatic in the treatment effect. From a Bayesian point of view, we only use the prior information orthogonal to the treatment effect. Hence, even if the investigator's prior is very informative about the treatment effect, the estimator will not reflect this ex-ante bias. The definition of investigator risk functions $\mathcal{R}^*$ as mean-squared error with respect to some pseudo-target therefore plays a second role. Alignment with respect to these preferences also implies robustness against misspecification of priors in the direction of the treatment effect. Hence, wrong preconceptions about treatment effects will not lead to systematic distortions in estimates if we restrict researchers to unbiased estimators.

5.3 Complete Class and Estimation-Prediction Duality

Since there is generally no single best estimator for all values of the truth, we have minimized average loss for some prior. If instead we consider admissible estimators that are not dominated by any other estimator in a purely frequentist sense, the same conclusions apply. Indeed, a duality result connects admissible unbiased estimation and admissible prediction.

For finite support any admissible estimator is the limit of a Bayes estimator that minimizes posterior loss given the data for some prior with full support (e.g. Ferguson, 1967). I extend this complete-class argument to unbiased estimators by applying it to the representation in Lemma 1.

Theorem 3 (Complete-class theorem for unbiased estimators)**.** *For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect*

to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ with full support such that $\hat{\tau}$ equals the limit of the respective estimators in Theorem 2. Conversely, for any converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ that put positive weight on every state $\theta \in \Theta$, the limit of the estimators is admissible among unbiased estimators.

The individual increments

$$\begin{aligned}\phi_i(z_{-i}) &= \mathrm{E}_\pi[\bar{y}_i|z_{-i}],\\ \phi_{i;j}(z_{-ij}) &= \mathrm{E}_\pi[\bar{\bar{y}}_i|z_{-ij}]\end{aligned}$$

solve a leave-one-out and leave-two-out out-of-sample prediction problem, respectively. (The adjustment $\phi_{ij}(z_{-ij})$ is obtained as $\phi_{ij}(z_{-ij}) = \phi_{i;j}(z_{-ij}) - \phi_{j;i}(z_{-ij})$.) Indeed, ϕ_i and $\phi_{i;j}$ minimize the average of the forecast risk

$$r^i_\theta(\hat{y}_i) = \mathrm{E}_\theta[w(d_i)(\hat{y}_i - y_i)^2] \tag{2}$$

given the respective data and the prior π. The weights

$$w(d_i) = \left(\frac{(d_i - p)}{p(1-p)}\right)^2, \qquad\qquad w(d_i) = \left(\frac{n(d_i n - n_1)}{n_1 n_0}\right)^2$$

put higher emphasis on the smaller of the the treatment and control groups.[19]

I apply the complete-class logic to both sides of the problem to obtain a one-to-many correspondence between unbiased admissible estimation and admissible prediction.[20] The relationship is not one-to-one because different prediction solutions may correspond to the same estimator.

Corollary 5.1 (Estimation-prediction duality)**.** *Any admissible unbiased estimator can be expressed in terms of a jointly admissible solution to the prediction problems with risks r^i_θ. Conversely, any jointly admissible solution to the prediction problems defined by risks r^i_θ yields an admissible unbiased estimator of the sample-average treatment effect via the representation in Lemma 1. (Here, by joint admissibility*

[19] This mirrors Lin's (2013) "tyranny of the minority" estimator, which puts similar weights into a least-squares regression.

[20] Wager et al. (2016) in an asymptotic framework using a similar sample-splitting construction note that "the precision of the treatment effect estimates obtained by such regression adjustments depends only on the prediction risk of the fitted regression adjustment." Similary, Wu and Gagnon-Bartsch (2017) show that the variance of their LOOP estimator is approximately the average mean-squared error in predicting the oracle adjustments, provided that certain covariance terms are negligible. In my finite-sample Bayesian setting, the duality holds exactly.

I mean that the solutions to all prediction problems are the limits of average-risk minimizers with respect to the same sequence of priors.)

While the estimator itself is unbiased, the implicit prediction solution of a low-variance estimator will typically have bias.

5.4 Constrained Cross-Fold Solutions

It may be infeasible to estimate all regression adjustments optimally. Mimicking machine-learning practice, one could instead partition the sample into K folds and estimate adjustments in one fold jointly from the units in all other folds. The resulting estimator resembles Wager et al.'s (2016) "cross-estimation" and Chernozhukov et al.'s (2017a) "cross-fitting" estimator.

Remark 5.1 (Exact K-fold cross-fitting)**.** *For a partition of the sample*

$$\{1,\dots,n\} = \bigcup_{k=1}^{K} \mathcal{I}^{(k)}$$

into K folds with $n^{(k)} \geq 2$ units each of which $n_1^{(k)} > 0$ treated and $n_0^{(k)} > 0$ untreated, the estimator

$$\hat{\tau}(z) = \frac{1}{n}\sum_{k=1}^{K} n^{(k)} \sum_{i\in\mathcal{I}^{(k)}} \frac{d_i n^{(k)} - n_1^{(k)}}{n_1^{(k)} n_0^{(k)}} \left(y_i - \phi_i^{(k)}(z_{-\mathcal{I}^{(k)}})\right)$$

is unbiased for the sample-average treatment effect τ conditional on $(\mathcal{I}^{(k)})_{k=1}^{K}$ and $(n_1^{(k)})_{k=1}^{K}$ under either randomization. The investigator obtains their constrained optimal (Bayes) $\hat{\tau}$ among these estimators at

$$\phi_i^{(k)}(z_{-\mathcal{I}^{(k)}}) = \mathrm{E}_\pi[n_0^{(k)} y_i(1) + n_1^{(k)} y_i(0) | z_{-\mathcal{I}^{(k)}}]/n^{(k)}.$$

Randomization could be within folds or folds could be chosen after overall randomization. If K divides n_1 and n_0, we achieve perfect balance by stratifying folds by treatment (or the other way around), $Kn_1^{(k)} = n_1$ and $Kn_0^{(k)} = n_0$.

In particular, the optimal regression adjustments are predictions even when not all adjustments are estimated. Indeed, $\phi_i^{(k)}$ minimizes average risk r_θ^i in (2) with

weight

$$w_i^{(k)}(d_i) = \left(\frac{n^{(k)}(d_i n^{(k)} - n_1^{(k)})}{n_1^{(k)} n_0^{(k)}} \right)^2$$

given data from other folds and the prior π. An unbiased estimator of the risk is the average loss on fold k.

5.5 Machine Learning Algorithms as Agents

When high-dimensional unit characteristics are available, machine learning offers a solution to the prediction problems implicit to unbiased estimation. Effectively, machine learning engages in automated specification searches to find a model that predicts well. I take a principal-agent perspective on machine-learning algorithms to provide a formal embedding. The investigator as principal delegates to the machine-learning agent. Through sample splitting, there is no misalignment of preferences between the investigator and the machine-learning agent provided the latter minimizes prediction risk, and the investigator achieves a second-best estimation solution from first-best predictions.

For randomly sampled units, the implicit prediction solutions forecast outcomes from characteristics. If units are draw according to the population distribution $(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P}$ that includes characteristics x_i, then

$$y_i(1), y_i(0) | x_1, \ldots, x_n \sim \mathrm{P}(x_i).$$

Increments $\phi_i(y_{\mathcal{T}_i}, d_{\mathcal{T}_i})$ fitted on $\mathcal{T}_i \subseteq \{1, \ldots, n\} \setminus \{i\}$ minimize expected forecast risk

$$\mathrm{E}[r_\theta^i(\hat{y}_i) | x_1, \ldots, x_n, y_{\mathcal{T}_i}, d_{\mathcal{T}_i}] = \mathrm{E}[\mathrm{E}_\theta[w(d_i)(\hat{y}_i - y_i)^2 | y_i(1), y_i(0)] | x_i]$$

over $\hat{y}_i \in \mathbb{R}$. Writing $\hat{y}_i = \hat{f}_i(x_i)$ with $\hat{f}_i : \mathcal{X} \to \mathbb{R}$ a function of training data $(y_{\mathcal{T}_i}, d_{\mathcal{T}_i}, x_{\mathcal{T}_i})$ evaluated on the test point x_i, $\hat{f}_i$ solves the prediction problem

$$L_i(\hat{f}) = \mathrm{E}[w(d_i)(\hat{f}(x_i) - y_i)^2 | x_i] \to \min_{\hat{f}}. \tag{3}$$

Here, I conflate the population distribution P with the sampling process to describe the distribution of observable data.

Supervised machine learning offers non-parametric solutions of out-of-sample prediction problems like (3) that are particularly suitable for high-dimensional charac-

teristics x_i. Since the test point (y_i, d_i, x_i) follows the same distribution as the training sample $\mathcal{T}_i$, sample-splitting techniques within the training sample allow for specification searches (in the form of model regularization and combination) to obtain good average predictions at the test point. Furthermore, the realized loss at i is an unbiased estimate of $L_i(\hat{f}_i)$.

I capture machine learning as an agent who minimizes average forecast risk for weighted loss $w(d_i)(\hat{f}(x_i) - y_i)^2$. The machine-learning agent's choice $\hat{f}_i$ may have complex structure that eludes causal interpretation and its parameters may not even be stable approximations of correlation patterns (Mullainathan and Spiess, 2017). However, the investigator as principal cares only about the forecast properties of the agent's solution.

Provided that the agent (approximately) minimizes risk, their choices are (approximately) aligned with the preferences of the investigator. There is no moral hazard from unobserved modeling decisions in the delegation of the prediction task from investigator to machine-learning agent. The machine-learning delegation task can be realized as a contract that pays the provider of the machine-learning solution according to the observed performance of prediction functions $\hat{f}_i$ on test points (y_i, d_i, x_i).

Crucially, sample splitting guards against prediction mistakes. Even when the specific prediction method does not minimize forecast risk or makes systematic mistakes, the resulting estimator is still unbiased. Worse predictions can lead to worse estimation performance, but only through variance.

6 Pre-Analysis Plans and Ex-Post Analysis

There are two ways in which we can guarantee that the investigator delivers an unbiased estimator. In the previous section, I derived a representation of unbiased estimators that require that the investigator's estimator only uses one part of the sample when constructing regression adjustments for another part. Since the investigator will ultimately work with all of the data, this condition cannot be verified ex-post, but has to be guaranteed by ex-ante commitment. One way to guarantee that the estimator fulfills this condition is to require that the investigator commits to the construction of all regression adjustments before she has seen any of the data.

In this section, I consider instead that the investigator commits to how she will split and distribute the data to one or multiple researchers who have not yet accessed

the data. Detailed commitment may be infeasible for methods that require active guidance by the researcher, impractical for very complex algorithms, or inefficient when some prior uncertainty is resolved only after the initial commitment. I therefore consider sample-splitting schemes that leave some or all regression adjustments unspecified, and instead delegate their estimation. Delegating to one researcher can already improve over simple pre-specified estimators. Delegating to two researchers attains semi-parametric efficiency without any commitment beyond sample splitting.

6.1 Automated vs Human Specification Searches

The results in this article imply a constructive characterization of robust yet flexible pre-analysis plans. The two ways of ensuring unbiasedness correspond to two different types of specification searches. The first way in which we can be flexible while also ensuring unbiasedness is that the investigator commits in her pre-analysis plan which algorithm she will use to construct regression adjustments. This algorithm then engages in automated specification searches to solve the prediction problems I have shown to be equivalent to unbiased estimation.

The second way in which specification searches remain possible applies when the investigator splits the sample and distributes it to one or multiple researchers. Then each researcher can search through specifications using his full subsample and does not have to commit to an empirical strategy ex ante. As long as the investigator commits to how she will distribute the sample and use the output from the researchers, and follows the procedures I characterize below, the resulting estimator is again guaranteed to be unbiased.

Automated and human specification searches can be combined to ensure precise and unbiased estimation under logistical constraints. An investigator who analyzes the data by herself can split the sample into two, apply a pre-specified algorithm to the first half of the data, and search through specifications by hand only in the second half.

6.2 Unbiased Estimators without Full Commitment

I show that the class of unbiased estimators includes protocols that do not require full pre-commitment, but leave additional degrees of freedom open. The investigator commits to an estimator that includes flexible inputs by one or multiple researchers. Each researcher obtains access to a subset of the data, but does not have to pre-

commit to their output.

Definition 2 (K-distribution contract)**.** *A K-distribution contract $\hat{\tau}^\Phi$ distributes data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ to K researchers. Researcher k receives data $g_k(z) \in A_k$ and returns the intermediate output $\hat{\phi}_k(g_k(z)) \in B_k$. The estimate is*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \Phi((\hat{\phi}_k(g_k(z))_{k=1}^K; z).$$

The investigator chooses the functions g_k (from data in $\mathcal{Z}$ to researcher input in A_k) and Φ (from the researcher outputs in $\bigtimes_{k=1}^K B_k$ and data in $\mathcal{Z}$ to estimates in $\mathbb{R}$) before accessing the data.

While the investigator still commits which part of the data individual researchers receive and how their choices and the data form an overall estimate, the individual researchers' actions are not pre-specified. From my results in the previous section, I obtain a full characterization of K-distribution contracts that are unbiased no matter the choices of the researchers. Since the resulting estimators are always unbiased, the preferences of the researchers, the investigator, and the designer over these contracts are aligned provided that the investigator and the researchers all minimize average risk for risk functions in $\mathcal{R}^*$ and have the same prior π.

Lemma 6.1 (Characterization of unbiased K-distribution contracts)**.** *A K-distribution contract $\hat{\tau}^\Phi$ is unbiased for the sample-average treatment effect τ_θ for any conformable researcher input $(\hat{\phi}_k)_{k=1}^K$ if and only if:*

1. *For known treatment probability p, there exist regression adjustments $(\phi_i : (\bigtimes_{k \in C_i} B_k) \times (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1-p)} (y_i - \phi_i((\hat{\phi}_k(g_k(z))_{k \in C_i}; z_{-i}))$$

   for $C_i = \{k; g_k(z) = \tilde{g}(z_{-i})$ for some $\tilde{g}\}$.

2. *For fixed number n_1 of treated units, there exist regression adjustments $(\phi_{ij} : (\bigtimes_{k \in C_{ij}} B_k) \times (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}((\hat{\phi}_k(g_k(z))_{k \in C_{ij}}; z_{-ij})),$$

   for $C_{ij} = \{k; g_k(z) = \tilde{g}(z_{-ij})$ for some $\tilde{g}\}$.

In other words, the regression adjustments of a given unit are only controlled by the choices of researchers who do not have access to data from that unit. The sets C_i, C_{ij} are thus the set of researchers who have control over regression adjustments ϕ_i, ϕ_{ij}. For the special case $K = 1$, this construction resembles proposals to use hold-out sets to avoid false positives in multiple testing (Dahl et al., 2008; Fafchamps and Labonne, 2016; Anderson and Magruder, 2017). For general K, the construction resembles K-fold cross-validation. Indeed, we obtain a particularly simple form if we restrict sample distribution to K-fold partitions.

Corollary 1 (Characterization of unbiased K-fold distribution contracts)**.** *For K disjoint folds $\mathcal{I}_k \subseteq \{1, \dots, n\}$ with projections $g_k : (y, d) = z \mapsto z_{-\mathcal{I}_k} = (y_i, d_i)_{i \neq \mathcal{I}_k}$, a K-distribution contract $\hat{\tau}^\Phi$ is unbiased if and only if:*

1. *For a known treatment probability p,there exist a fixed unbiased estimator $\hat{\tau}_0(z)$ and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z_{-i})$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

2. *For a fixed number n_1 of treated units, there exist a fixed unbiased estimator $\hat{\tau}_0(z)$ and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^{K} \sum_{\{i<j\} \subseteq \mathcal{I}_k} (d_i - d_j) \phi_{ij}^k(z_{-ij}),$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

K-fold distribution contracts are similar to K-fold cross-fitting from Remark 5.1, but different in terms of motivation and more flexible in terms of application. K-fold distribution is motivated by ensuring unbiasedness, not by computational limitations. While K-fold cross-fitting is contained as the special case where a researcher determines the regression adjustments for all units in the target fold directly from their training data (that is, no data from the target fold is used to adjust any of the units in that fold), K-fold distribution contracts also contain solutions that use additional data without bias. Indeed, for the case of known p, say, if regression adjustments take the form $\phi_i^k(\lambda_k; z_{-i})$ with a pre-determined function $\phi_i^k(\cdot; \cdot)$ and

some tuning parameter λ_k, then the adjustments can be be a function of all the data in z_{-i} as long as the tuning parameter λ_k is fitted only on the other folds.[21]

6.3 Hybrid Pre-Analysis Plans

I apply the previous result to show that a simple pre-analysis plan is dominated by a hybrid pre-analysis plan that allows for additional discretion after part of the data is revealed. The investigator fixes some regression adjustment, but can modify others after access to a subset of the sample. Since sample splitting ensures preference alignment, the hybrid estimator will dominate if the ex-post analysis permits better implementation of prior information.

I now assume that the investigator's prior π is only realized after the data is available. Before the data is available, the investigator has a prior η^I over π. I think of η^I as a crude approximation to π. A simple ex-ante prior η^I could come from high costs of fully writing down or automating the way in which the investigator translates prior information and data into predictions of potential outcomes. The ex-post prior π could also represent updated beliefs after the pre-analysis plan has been filed. In both cases, however, the difference does not represent the information in the collected data itself, which will be incorporated in the posterior distribution instead.

Anderson and Magruder (2017) propose a hybrid pre-analysis plan for multiple testing. The investigator pre-specifies some hypothesis they will test, and then selects additional hypotheses from a training sample. The additional hypotheses are only evaluated on the remaining hold-out sample. I adopt their proposal to my estimation setting.

Definition 6.1 (Hybrid pre-analysis plan)**.** *A hybrid pre-analysis plan is a 1-fold distribution contract, i.e. an estimator*

$$\hat{\tau}^{\Phi}(\hat{\phi}; z) = \Phi(\hat{\phi}(z_{\mathcal{T}}); z)$$

that pre-specifies a mapping Φ from ex-post researcher input $\hat{\phi}(z_{\mathcal{T}})$ and realized sample data z to an estimate of the sample-average treatment effect. The researcher

[21] This idea can be applied to the post-LASSO (Belloni et al., 2013) after selection on the training sample. Unlike the cross-fitted LASSO, the post-selection fitting step can include the full sample (provided all regression adjustments are fitted using a leave-one- or leave-two-out construction). Furthermore, the selection step can include researcher intervention that has not been pre-specified.

(which here could be the investigator herself) obtains access to training data $\mathcal{T} \subseteq \{1, \ldots, n\}$ *before the final estimator is formed.*

I assume that the investigator must still pre-commit to an unbiased estimator, so Corollary 1 for $K = 1$ fully characterizes the plans available to the investigator. In these sample-splitting plans, the choices of the researcher after gaining access to the training sample are fully aligned with the intentions of the investigator according to their updated prior. The investigator pre-commits all adjustments in the training sample according to η^I, while the researcher chooses the remaining regression adjustments according to π and their training data.

Theorem 6.1 (Hybrid pre-analysis plan dominates rigid pre-analysis plan)**.** *Assume that investigator and researcher have risk functions in* $\mathcal{R}^*$*. The optimal unbiased pre-committed estimator* $\hat{\tau}^{\text{pre}}$ *is strictly dominated by an unbiased hybrid pre-analysis plan with respect to average variance, i.e. the hybrid plan is as least as precise on average over any ex-ante prior* η^I *and strictly better for many non-trivial ex-ante priors* η^I*.*

Since the researcher's and investigator's preference over unbiased estimators is fully aligned with the designer's goal, there is no preference misalignment and the variance captures all of their risk functions.

Remark 6.1 (Optimal hybrid pre-analysis plan)**.** *The dominating hybrid plan is:*

1. *For known treatment probability* p*, the researcher chooses regression adjustments* $(\phi_i^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i \notin \mathcal{T}} = \hat{\phi}(z_{\mathcal{T}})$ *to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n} \sum_{i \notin \mathcal{T}} \frac{d_i - p}{p(1-p)} \, \phi_i^{\text{post}}(z_{-i})$$

   where $1 \leq |\mathcal{T}| \leq n - 1$*.*

2. *For fixed number* n_1 *of treated units, the researcher chooses adjustments* $(\phi_{ij}^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{\{i<j\} \cap \mathcal{T} = \emptyset} = \hat{\phi}(z_{\mathcal{T}})$ *to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n_1 n_0} \sum_{\{i<j\} \cap \mathcal{T} = \emptyset} (d_i - d_j) \, \phi_{ij}^{\text{post}}(z_{-ij})$$

   where $1 \leq |\mathcal{T}| \leq n - 2$*.*

In both cases, the investigator commits to the training sample $\mathcal{T} \subseteq \{1, \ldots, n\}$ *and the unbiased estimator* $\hat{\tau}^{\text{pre}} : \mathcal{Z} \to \mathbb{R}$.

The optimal ex-post adjustments modify the implicit adjustments of the ex-ante estimator to match the solution from Theorem 2 on the relevant subset, i.e. they solve an out-of-sample prediction problem.

6.4 Many-Researcher Delegation

The hybrid pre-analysis plan is itself dominated by a plan that distributes the data to multiple researchers. If a single researcher has access to the full dataset before committing their estimator, bias can return even if the researcher represents their estimate by regression adjustments. Distribution to multiple researchers reduces inefficiency without introducing misalignment. Even when ex-ante commitment beyond a trivial estimator is infeasible or undesirable, distribution between at least two researchers can produce an ex-post desirable estimator.

Remark 6.2. *Assume that the investigator and researchers all have risk functions in* $\mathcal{R}^*$*, and that the researchers all share the same (ex-post) prior* π*. Then an optimal unbiased* K*-distribution contract is dominated by an unbiased* $K + 1$*-distribution contract in the sense of Theorem 6.1.*

I now consider standard large-sample efficiency criteria for the estimation of the population-average treatment effect. There is no unique variance-minimal solution in finite samples, as the class of admissible estimators is large. In the large-sample limit, however, essentially all admissible estimators have approximately equal performance, and coordination between researchers with different (non-dogmatic) priors is resolved by a common understanding of the truth.

Under random sampling of units, the semi-parametric efficiency bound of Hahn (1998) is achieved at oracle prediction adjustments.[22] For $(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \text{P}$ with fixed probability p of treatment, an infeasible estimator of the population average treatment effect τ is

$$\hat{\tau}^{\text{P}}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \text{E}[\bar{y}_i | x_i])$$

[22] See also Imbens (2004) for a discussion of efficient estimation of average treatment effects.

where the oracle regression adjustments are optimal given knowledge of P. While we will not generally be able to achieve the variance of $\hat{\tau}^{\mathrm{P}}$, under assumptions we can achieve a variance that is asymptotically equivalent (i.e. $\mathrm{Var}(\hat{\tau})/\mathrm{Var}(\hat{\tau}^{\mathrm{P}}) \to 1$ as $n \to \infty$).

Remark 6.3 (Semi-parametric efficiency)**.** *If researchers use prediction algorithms* $(A_n : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}, z \mapsto \hat{f}_n)_{n=1}^{\infty}$ *with*

$$\mathrm{E}[(\hat{f}_n(x_i) - \mathrm{E}[\bar{y}_i|x_i])^2] \to 0$$

as $n \to \infty$*, then delegation to two researchers with risk functions in* $\mathcal{R}^*$ *(who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased estimation of* τ_θ*, and large-sample semi-parametric efficient estimation of* τ *for the semi-parametric efficiency bound of Hahn (1998).*

In other words, semi-parametric efficiency is achieved from distribution of the data to at least two independent researchers with risk-consistent predictors. Data distribution ensures that there is no misalignment.

Conclusion

By taking a mechanism-design approach to econometrics, I account for misaligned researcher incentives in causal inference. I motivate why and how we should pre-commit our empirical strategies, and demonstrate that there exist flexible pre-analysis plans that allow for exploratory data analysis and machine learning without leaving room for biases. In particular, I characterize all unbiased estimators of an average treatment effect as sample-splitting procedures that permit beneficial specification searches.

My results shed light on the role of bias and variance in treatment-effect estimation from experimental data. Allowing for bias can reduce the variance and thus improve precision. But when incentives are misaligned, giving a researcher the freedom to choose the bias may, in fact, reduce precision. However, once we restrict the researcher to unbiased estimators, there will again be a bias-variance tradeoff in the nuisance parameters associated with the control variables. I have shown in this article that unbiased estimation of a treatment effect in an experiment is equivalent to a set or prediction tasks. Inside these tasks, some bias in return for a substantial variance reduction can improve prediction quality. Better predictions in turn

translate into lower variance of the unbiased estimator.

In related work, I show that under additional parametric assumptions standard treatment-effect estimators are dominated because shrinkage can reduce variance without introducing bias. In a linear model with homoscedastic, Normal noise and exogenous treatment, the usual linear least-squares estimator for the treatment effect is dominated provided that there are at least three Normally-distributed control variables (Spiess, 2017b).[23] In that case, I reduce variance without introducing bias by James and Stein (1961) shrinkage in the underlying prediction problem.[24]

I am working on extending my mechanism-design approach to other estimation tasks in experimental or quasi-experimental data. Applications include effects on endogenously chosen subgroups, heterogeneous treatment effects, treatment effects under optimal assignment, and tests for effects on multiple outcome variables. In each case, I conjecture that my approach can motivate a design restriction by its preference alignment property, yield a representation of the resulting estimators as sample-splitting procedures, and suggest a characterization of optimal mechanisms and second-best pre-analysis plans.

One possible direction to pursue is to extend the approach of this article to cases where unbiased estimators are generally unavailable. In instrumental-variable estimation, unbiased estimation is possible under sign restrictions on the first stage (Andrews and Armstrong, 2017), but generally infeasible when the parameter space is unrestricted (Hirano and Porter, 2015). Still, when there are many instruments, we can improve estimation by providing better solutions to the first-stage prediction problem implicit to the two-stage linear IV model. For example, shrinkage in the first stage reduces bias relative to the standard two-stage least-squares estimator (Spiess, 2017a). This finding raises the question how the delegation of the first-stage prediction problem can be realized in a way that aligns researcher preferences.

[23]The usual linear least-squares estimator is, by Gauss-Markov, still variance-minimal among conditionally unbiased estimators. However, once we integrate over the distribution of control variables (and if these are orthogonal to treatment), I show that there is an unbiased estimator with lower variance.

[24]In the nonparametric setting in this paper and the Normal-linear setting in Spiess (2017b), unbiased estimation reduces to prediction problems. The results are connected because they both stem from invariances that characterize the distributions - in the case of this paper reflections and permutations, in the case of Spiess (2017b) rotations that leave the Normal distribution invariant.

References

Anderson, M. L. and Magruder, J. (2017). Split-sample strategies for avoiding false discoveries. *NBER working paper*.

Andrews, I. and Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, 8(2):479–503.

Andrews, I. and Kasy, M. (2017). Identification of and correction for publication bias. *NBER working paper*.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Aronow, P. M. and Middleton, J. A. (2013). A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments. *Journal of Causal Inference*, 1(1).

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Balzer, L. B., van der Laan, M. J., Petersen, M. L., and the SEARCH Collaboration (2016). Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in Medicine*, 35(25):4528–4545.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Carroll, G. (2015). Robustness and linear contracts. *The American Economic Review*, 105(2):536–563.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017a). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017b). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Cochran, W. G. (1972). Observational studies. *Reprinted in Observational Studies, 2015*.

Coffman, L. C. and Niederle, M. (2015). Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible. *Journal of Economic Perspectives*, 29(3):81–98.

Dahl, F. A., Grotle, M., Šaltytė Benth, J., and Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. *European Journal of Epidemiology*, 23(4):237–242.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Fafchamps, M. and Labonne, J. (2016). Using Split Samples to Improve Inference about Causal Effects. *NBER working paper*.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Academic press.

Frankel, A. (2014). Aligned Delegation. *American Economic Review*, 104(1):66–83.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Glaeser, E. L. (2006). Researcher incentives and empirical methods. *NBER working paper*.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315.

Hájek, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, pages 1124–1147.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Heckman, J. J. and Singer, B. (2017). Abducting Economics. *American Economic Review*, 107(5):298–302.

Hirano, K. and Porter, J. R. (2015). Location properties of point estimators in linear instrumental variables and related models. *Econometric Reviews*, 34(6-10):720–733.

Hirano, K. and Wright, J. H. (2017). Forecasting With Model Uncertainty: Representations and Risk Reduction. *Econometrica*, 85(2):617–643.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945.

Holmström, B. R. (1978). *On Incentives and Control in Organizations*. PhD thesis, Stanford University.

Holmström, B. R. (1984). On the theory of delegation. *Bayesian models in economic theory*.

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

Hurwicz, L. and Shapiro, L. (1978). Incentive structures maximizing residual gain under incomplete information. *The Bell Journal of Economics*, pages 180–191.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Fourth Berkeley Symposium*.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(03):324–338.

Leamer, E. E. (1974). False models and post-data model construction. *Journal of the American Statistical Association*, 69(345):122–131.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. Wiley.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

Lenz, G. and Sahn, A. (2017). Achieving statistical significance with covariates.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Translated in Statistical Science, 1990*.

Olken, B. A. (2015). Promises and Perils of Pre-Analysis Plans. *Journal of Economic Perspectives*, 29(3):61–80.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–aac4716.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, pages 233–239.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58.

Spiess, J. (2017a). Bias Reduction in Instrumental Variable Estimation through First-Stage Shrinkage. *arXiv preprint arXiv:1708.06443*.

Spiess, J. (2017b). Unbiased Shrinkage Estimation. *arXiv preprint arXiv:1708.06436*.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa. *Journal of the American Statistical Association*, 54(285):30.

Tullock, G. (1959). Publication decisions and tests of significance—a comment. *Journal of the American Statistical Association*, 54(287):593–593.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

Wald, A. (1950). *Statistical decision functions.* Wiley.

Wu, E. and Gagnon-Bartsch, J. (2017). The loop estimator: Adjusting for covariates in randomized experiments. *arXiv preprint arXiv:1708.01229*.

Appendix

This appendix builds up to the proofs of the main results (Theorem 1, Lemma 1, Theorem 2). Additional proofs and results I collect in the Supplementary Appendix. Throughout, I restate the relevant claims from the main paper with their original numbering. I prepend the letter of the respective section to additional and auxiliary results.

A Minimax Optimality of Fixed Bias

Lemma 4.1 (Unbiasedness aligns estimation)**.** *If the investigator has risk from $\mathcal{R}^*$ then the investigator will choose from the unbiased estimators $\mathcal{C}^*$ according to the designer's preferences.*

Proof of Lemma 4.1. Take any investigator risk function $r^I \in \mathcal{R}^*$, unbiased estimator $\hat{\tau} \in \mathcal{C}^*$, and prior $\pi \in \Delta(\Theta)$. ($\Delta(\Theta)$ denotes the unit $|\theta| - 1$-simplex in $\mathbb{R}^\Theta$.) Then, the designer's average risk is

$$\begin{aligned}
&\mathrm{E}_\pi[r^D_\theta(\hat{\tau})] \\
\overset{r^I \in \mathbb{R}^*}{=}\ &\mathrm{E}_\pi[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \\
=\ &\mathrm{E}_\pi[((\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)]) - (\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta))^2] \\
=\ &\mathrm{E}_\pi[(\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)])^2] + \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta)^2] \\
\overset{\hat{\tau} \in \mathcal{C}^*}{=}\ &\mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))] + \mathrm{E}_\pi[(\tau_\theta - \tilde{\tau}_\theta)^2]
\end{aligned}$$

by a bias-variance decomposition. (I conflate P_θ into P_π.) Since $\mathrm{E}_\pi[(\tau_\theta - \tilde{\tau}_\theta)^2]$ is constant with respect to $\hat{\tau}$ and $\mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))]$ does not vary with $\tilde{\tau}$, the estimation target $\tilde{\tau}$ does not affect the choice of the estimator from $\mathcal{C}^*$. Hence, choices are as if $\tilde{\tau} = \tau$. The investigator chooses from $\mathcal{C}^*$ according to the designer's risk r^D. □

Theorem 1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1, i.e.

$$\mathcal{C}^{\eta} \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_{\eta}\left[r_{\eta}^{D}\left(\arg\min_{\hat{\tau}\in\mathcal{C}} \mathrm{E}_{\pi}[r_{\theta}^{I}(\hat{\tau})]\right)\right].$$

Proof of Theorem 1. I apply the strategy from Theorem 1 in Frankel (2014) to establish that the unbiasedness restriction yields a minimax (maxmin in utility terms) optimal mechanism. Relative to the quadratic-loss constant-bias setup in Frankel (2014), average risk yields weighted sums where the prior changes weights and the bias changes across decisions (sample draws) and states (posterior expectations). Rather than using Lemma 3 on quadratic-loss constant-bias utilities in Frankel (2014) as stated there, I therefore appeal directly to the logic of his more general Theorem 1, which I extend to deal with the non-compact type and action spaces in my application.

The agent's (investigator's) actions are the estimates $\hat{\tau}(z)$ at all $N = (2|\mathcal{Y}|)^n$ sample points $z \in \mathcal{Z}$. (I assume that the covariates x are already known when the investigator commits to their estimator.) The state that only the agent observes is the investigator's prior $\pi \in \Delta(\Theta)$. π is drawn from the (hyper-)prior η.

In the parlance of Frankel (2014), I consider the Φ-moment mechanisms where the agent chooses from estimators

$$\mathcal{C}_{\beta} = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_{\theta}[\hat{\tau}] = \tau_{\theta} + \beta_{\theta} \forall \theta \in \Theta\}$$

for a set of fixed biases $\beta \in \mathbb{R}^{\Theta}$. (Each expectation – a weighted sum over actions $\hat{\tau}(z)$ – is a map from actions to real numbers.) To show that this mechanism is maxmin optimal for some choice of β, I establish that:

1. Any feasible such Φ-moment mechanism (i.e. any bias vector β with $\mathcal{C}_{\beta} \neq \emptyset$) induces aligned delegation over $\mathcal{R}^*$, that is, subject to the restriction $\hat{\tau} \in \mathcal{C}_{\beta}$ agents of all risk types $r^I \in \mathcal{R}^*$ choose as if they were of risk type r^D.

2. $\mathcal{R}^*$ is Φ-rich, that is, for any mechanism there exists some $\overline{\beta} \in \mathbb{R}^{\Theta}$ and a sequence of risk types $(r^{I_k})_{k=1}^{\infty} \in (\mathcal{R}^*)^{\mathbb{N}}$ such that for all realized $\pi \in \Delta^*(\Theta)$ and all corresponding sequences $(\hat{\tau}_k)_{k=1}^{\infty}$ of chosen estimators, $\lim_{k\to\infty} \mathrm{E}_{\theta}[\hat{\tau}_k(z)] = \tau_{\theta} + \overline{\beta}_{\theta}$ for all θ in the support of π. (Unlike Frankel (2014) I do not explicitly consider mixed strategies since randomized estimators are dominated in my setting.)

Similar to Frankel's (2014) Theorem 1, the restriction $\mathcal{C}_\beta$ is then minimax optimal provided that β is chosen to minimize the designer's average risk, for some distribution (hyperprior) η over π. I will develop this deduction below for my specific case (in which type and action spaces are not compact) once I have established aligned delegation and richness.

1. Aligned delegation. For $\beta \in \mathbb{R}^\Theta$ such that $\mathcal{C}_\beta \neq \emptyset$, the average over risk $r^I \in \mathcal{R}^*$ for an estimator $\hat{\tau} \in \mathcal{C}_\beta$ over the prior $\pi \in \Delta(\Theta)$ is

$$\mathrm{E}_\pi r^I_\theta(\hat{\tau}) = \mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))] + \mathrm{E}_\pi[(\tau_\theta + \beta_\theta - \tilde{\tau}_\theta)^2]$$

as in the proof of Lemma 4.1. Hence, choices do not vary with the risk type of the investigator and are as if the investigator shared the designer's risk function r^D.

2. Richness. For some arbitrary, but fixed mechanism, our goal is to find a vector of biases $\bar{\beta}$ and a risk sequence $r^{I_1}, r^{I_2}, \ldots$ such that biases of mechanism outcomes along this sequence always converge to $\bar{\beta}$. I first justify assumptions on the mechanism, then pick a bias vector $\bar{\beta}$, and finally construct a suitable sequence of risk types that ensures bias convergence.

For some conformal mechanism, consider the set $\mathcal{C} \subseteq \mathbb{R}^\mathcal{Z}$ of estimators $\hat{\tau}$ that are outcomes for some investigator risk function $r^I \in \mathcal{R}^*$ and prior π in the support of η. Note that the outcomes of the mechanism are the investigator choices

$$\hat{\tau}_\pi(r^I) \in \operatorname*{arg\,min}_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi r^I_\theta(\hat{\tau}) \tag{4}$$

where by assumption ties are broken in favor of the designer. I first show that $\mathcal{C}$ in (4) is wlog closed. Since the minimizers are already included in $\mathcal{C}$, taking the closure of $\mathcal{C}$ does not change investigator risk at their optimal choices. Replacing $\mathcal{C}$ by its closure thus does not affect investigator risk at choices (4), and can only improve outcomes for the designer, since additional ties are broken in their favor. For the analysis of minimax optimal mechanisms, we can therefore assume wlog that $\mathcal{C}$ is closed.

I first assume that $\mathcal{C}$ is also bounded. Define the set

$$\mathcal{D} = \{\theta \mapsto \mathrm{E}_\theta[\hat{\tau}(z)]; \hat{\tau} \in \mathcal{C}\} \subseteq \mathbb{R}^\Theta$$

of vectors of expectations achieved by estimators in $\mathcal{C}$. By linearity of expectation, $\mathcal{D}$ is wlog compact by the above reasoning. Fix some ordering $\theta_1, \ldots, \theta_J$ of Θ (where $J = |\Theta|$). Let δ^0 be the maximal element in $\mathcal{D}$ with respect to the corresponding lexicographic ordering (so that, in particular, $\delta^0_{\theta_1} \geq \delta_{\theta_1}$ for all $\delta \in \mathcal{D}$). For every $h \in \{2, \ldots, J\}$, there exists a function $f_h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that for all $\varepsilon > 0$

$$\delta \in \mathcal{D}, \sum_{j=1}^{h-1} |\delta_{\theta_j} - \delta^0_{\theta_j}| < f_h(\varepsilon) \qquad \Rightarrow \qquad \delta_{\theta_h} < \delta^0_{\theta_h} + \varepsilon. \tag{5}$$

Indeed, assume not, then there must be some h and some $\varepsilon > 0$ such that for every $k \in \mathbb{N}$ there exists a $\delta^k \in \mathcal{D}$ with $\sum_{j=1}^{h-1} |\delta^\nu_{\theta_j}| < 1/k$ and $\delta^k_{\theta_h} \geq \delta^0_{\theta_j} + \varepsilon$. Since $\mathcal{D}$ is compact, δ^k must have a convergent subsequence with limit $\delta^\varepsilon \in \mathcal{D}$. But $\delta^\varepsilon_{\theta_j} = \delta^0_{\theta_j}$ for $j < h$ and $\delta^\varepsilon_{\theta_h} \geq \delta^0_{\theta_h} + \varepsilon > \delta^0_{\theta_h}$, contradicting that δ^0 is maximal in $\mathcal{D}$ with respect to the lexicographic order. Hence there exists such f_h, and we can assume wlog $\frac{f_h(\varepsilon)}{\varepsilon}$ is monotonically increasing in $\varepsilon > 0$ (otherwise we can choose an f_h that is smaller for small values of ε).

Given the target $\delta^0 \in \mathcal{D}$ and the functions $f_h, h \geq 2$, I construct a sequence of risk functions r^{I_k} such that the expectation of the corresponding investigator choices converges to δ^0 for all $\pi \in \Delta^*(\Theta)$. Concretely, for $k \in \mathbb{N}$ define $\alpha^k \in \mathbb{R}^\Theta$ recursively by

$$\alpha^k_{\theta_J} = k \qquad\qquad \alpha^k_{\theta_j} = k / \min_{h>j} f_h(1/\alpha^k_{\theta_h}), j < J$$

and consider the sequence of investigator risk functions

$$r^{I_k}_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}^k_\theta)^2], \qquad\qquad \tilde{\tau}^k_{\theta_j} = \delta^0_{\theta_j} + \alpha^k_{\theta_j}$$

which falls within $\mathcal{R}^*$.

For the case of bounded $\mathcal{C}$ and some arbitrary, but fixed $\pi \in \Delta^*(\Theta)$, it remains to show that the expectation of $\hat{\tau}_\pi(r^{I_k})$ converges to δ^0. Write $\delta^k_\theta = \mathrm{E}_\theta \hat{\tau}_\pi(r^{I_k})$. Assume for contradiction that δ^k_θ does not converge to δ^0_θ. Since also $\delta^k_\theta \in \mathcal{D}$ for all k and $\mathcal{D}$ compact, $(\delta^k_\theta)_{k=1}^\infty$ must have a converging subsequence $(\delta^{k_\ell}_\theta)_{h=1}^\infty$ with $\delta^{k_\ell} \to \delta^1 \in \mathcal{D} \setminus \{\delta^0\}$ as $h \to \infty$. The average investigator loss along the sequence is

$$\mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}})) = \mathrm{E}_\pi \underbrace{\mathrm{Var}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}}))}_{\leq \text{const. } (\mathcal{C} \text{ bounded})} + \mathrm{E}_\pi(\delta^{k_\ell}_\theta - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2. \tag{6}$$

Note that an estimator $\hat{\tau}^0$ with expectation $\delta^0 \in \mathcal{D}$ would also have been available in $\mathcal{C}$ by definition of $\mathcal{D}$, and the difference in risk between the chosen subsequence and the alternative is

$$\begin{aligned}
\Delta_\ell &= \mathrm{E}_\pi r_\theta^{I_{k_\ell}}(\hat{\tau}_\pi(r^{I_{k_\ell}})) - \mathrm{E}_\pi r_\theta^{I_{k_\ell}}(\hat{\tau}^0) \\
&\overset{(6)}{=} \mathrm{E}_\pi(\delta_\theta^{k_\ell} - (\delta_\theta^0 + \alpha_\theta^{k_\ell}))^2 - \mathrm{E}_\pi(\alpha_\theta^{k_\ell})^2 + \mathcal{O}(1) \\
&= \mathrm{E}_\pi \underbrace{(\delta_\theta^{k_\ell} - \delta_\theta^0)^2}_{\to(\delta_\theta^1 - \delta_\theta^0)^2} - 2\,\mathrm{E}_\pi(\delta_\theta^{k_\ell} - \delta_\theta^0)\alpha_\theta^{k_\ell} + \mathcal{O}(1) \\
&= -2\sum_{j=1}^{J} \pi(\theta_j)\alpha_{\theta_j}^{k_\ell}(\delta_{\theta_j}^{k_\ell} - \delta_{\theta_j}^0) + \mathcal{O}(1).
\end{aligned}$$

Denote by h the smallest index of for which $\delta_{\theta_h}^0 \neq \delta_{\theta_h}^1$. Since δ^0 is maximal with respect to the lexicographic ordering of $\mathcal{D}$ and δ^1 also in $\mathcal{D}$, we must have $\delta_{\theta_h}^0 - \delta_{\theta_h}^1 > 0$. By revealed preference and since $\alpha_{\theta_{j+1}}^k = o(\alpha_{\theta_j}^k)$ for all j, it follows that

$$0 \geq \Delta_\ell / \alpha_{\theta_h}^{k_\ell} = -2\sum_{j=1}^{h-1} \pi(\theta_j)\frac{\alpha_{\theta_j}^{k_\ell}}{\alpha_{\theta_h}^{k_\ell}}(\delta_{\theta_j}^{k_\ell} - \delta_{\theta_j}^0) - 2\pi(\theta_h)(\delta_{\theta_h}^1 - \delta_{\theta_h}^0) + o(1).$$

In particular, for $\varepsilon = \pi(\theta_h)(\delta_{\theta_h}^0 - \delta_{\theta_h}^1)$,

$$\liminf_{\ell\to\infty} \sum_{j=1}^{h-1} \underbrace{\pi(\theta_j)\frac{\alpha_{\theta_j}^{k_\ell}}{\alpha_{\theta_h}^{k_\ell}}(\delta_{\theta_j}^{k_\ell} - \delta_{\theta_j}^0)}_{=a_j^\ell} \geq \varepsilon > 0. \tag{7}$$

Hence there must exists some h^* and a subsequence ℓ_s such that

$$a_{h^*}^{\ell_s} \to \nu \in (0, \infty], \qquad \limsup_{s\to\infty} \frac{a_j^{\ell_s}}{a_{h^*}^{\ell_s}} \leq 1 \ \forall j < h. \tag{8}$$

(That is, $a_{h^*}^{\ell_s}$ is a maximal sequence within that subsequence, for a suitable asymptotic notion of maximality; it is not unique, but an instance can be constructed from iterated subsequences.) For simplicity, I write $k_s = k_{\ell_s}$. I assume wlog that

$\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_j} > 0$ for all s. By (6),

$$\sum_{j=1}^{h^*-1} |\delta^{k_s}_{\theta_j} - \delta^0_{\theta_j}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}}),$$

so there must exist some $j^* < h^*$ and a refinement of the subsequence along which $|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})/(h^*-1)$. Note that

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^*-1)}\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}}\frac{f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})}{\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}}}.$$

By (8) there exists some $\nu_0 \in (0,\infty)$ such that $a^{\ell_s}_{h^*} \geq \nu_0$ for all large s. By the definition of $a^{\ell}_{h^*}$ we find, again for large s, that

$$\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}} = \frac{a^{\ell_s}_{h^*}}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}} \geq \frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}.$$

By monotonicty of $\frac{f_{h^*}(\varepsilon)}{\varepsilon}$ therefore for large s

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^*-1)}\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}}\frac{f_{h^*}\left(\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}\right)}{\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}}.$$

By construction of the rates α^k_θ, we have that for every triple $j^* < h^* < h$ and every constant $c > 0$ and all large k

$$\begin{aligned}\frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_h}} f_{h^*}\left(c\frac{\alpha^k_{\theta_h}}{\alpha^k_{\theta_{h^*}}}\right) &\geq \frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_J}} f_{h^*}\left(c\frac{\alpha^k_{\theta_J}}{\alpha^k_{\theta_{h^*}}}\right) = \frac{\alpha^k_{\theta_{j^*}}}{k} f_{h^*}\left(\frac{ck}{\alpha^k_{\theta_{h^*}}}\right) \\ &\geq c\alpha^k_{\theta_{j^*}} f_{h^*}\left(\frac{1}{\alpha^k_{\theta_{h^*}}}\right) \geq ck \to \infty.\end{aligned}$$

It follows that

$$\frac{\pi(\theta_{j^*})\frac{\alpha_{\theta_{j^*}}^{k_s}}{\alpha_{\theta_h}^{k_s}}|\delta_{\theta_{j^*}}^{k_s}-\delta_{\theta_{j^*}}^{0}|}{\pi(\theta_{h^*})\frac{\alpha_{\theta_{h^*}}^{k_s}}{\alpha_{\theta_h}^{k_s}}(\delta_{\theta_{h^*}}^{k_s}-\delta_{\theta_{h^*}}^{0})}\to\infty.$$

By (8), $\delta_{\theta_{j^*}}^{k_s}-\delta_{\theta_{j^*}}^{0}<0$ for all but at most finitely many s. Hence $a_{j^*}^{\ell_s}/a_{h^*}^{\ell_s}\to-\infty$, and thus $\sum_{j=1}^{h-1}a_j^{\ell_s}\to-\infty$, contradicting (7). Therefore $\delta^1=\delta^0$.

Consider now the case when $\mathcal{C}$ is unbounded. First, if $\mathcal{C}$ is unbounded but $\mathcal{B}$ is still bounded (and thus wlog compact by linearity of the expectation projection), then the same argument as above goes through since there is always an estimator with finite variance and expectation δ^0 available (and the investigator minimizes variance given expectation), so unbounded variance along the investigator path can only make the choice with expectation δ^0 more attractive.

Second, if $\mathcal{B}$ is also unbounded, then $\mathcal{C}$ cannot be minimax optimal. Since $\mathcal{B}$ is unbounded, it must contain a sequence $\delta^k\in\mathcal{B}$ with $\|\delta^k\|$ diverging. The projection of δ^k on the unit sphere towards the origin must contain a converging subsequence with limit v where $\|v\|=1$. Consider a sequence of investigators with $\tilde{\tau}^k=v$ along the ray defined by the direction of this cluster point. One, if the average variance along the sequence of investigator choices is unbounded, then so is the average risk of the designer. Two, if the average variance along the sequence of investigator choices is bounded, then the bias diverges and average risk of the designer is again unbounded. Indeed, I show that it is not possible that both average variance and average expectation remain bounded along the ray. If the expectation vector $\mathrm{E}_\theta[\hat{\tau}(z)]$ along that sequence of investigators remains bounded, pick a point arbitrarily close to the ray that falls outside that bound. (Such a point exists by construction of v.) As investigator preference moves along the ray, the gain in average investigator risk from moving to that point outweigh any cost in terms of variance since the marginal cost of being off the expectation target only increases, while the variance cost remains bounded. Hence, the bias cannot remain bounded and the average risk of the designer diverges.

We therefore have that for any $\pi\in\Delta^*(\Theta)$ the bias of investigator choices along the sequence r^{I_k} converges to $\bar{\beta}_\theta=\delta_\theta^0-\tau_\theta$ for all $\theta\in\Theta$.

Proof of minimax optimality. Given any mechanism, by richness there exists a sequence of investigator risk functions r^{I_k} in $\mathcal{R}^*$ and a bias vector $\bar{\beta}$ such that $\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta \to \bar{\beta}_\theta$ for all $\pi \in \Delta^*(\Theta)$ and all $\theta \in \Theta$. The expected average designer's risk along this sequence is

$$\mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] = \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \underbrace{(\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta)^2}_{\to \bar{\beta}_\theta^2 \forall \theta \in \Theta, \pi \in \Delta^*(\Theta)},$$

where I omit the argument z of the estimators. Since biases are bounded (since $\mathcal{D}$ is) and the support of η is in $\Delta^*(\Theta)$, by dominated convergence

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] &= \liminf_{k\to\infty} \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}_\theta^2 \\ &\geq \mathrm{E}_\eta \liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}_\theta^2.\end{aligned}$$

For fixed $\pi \in \Delta^*(\Theta)$, $\liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k}))$ is at least the minimal asymptotic variance along a sequence $\hat{\tau}_\pi^k$ with bounded bias that converges to $\bar{\beta}$, and is otherwise unrestricted. Take such a sequence for which $\mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi^k)$ converges to its minimal limit. Along this sequence, $\hat{\tau}_\pi^k$ must be bounded, so it must have a convergent subsequence with some limit $\hat{\tau}_\pi^0$ in $\mathbb{R}^{\mathcal{Z}}$ for which by continuity also $\mathrm{E}_\theta[\hat{\tau}_\pi^0] - \tau_\theta = \bar{\beta}_\theta$. But then the variance of $\hat{\tau}_\pi^0$ must be at least the variance of a variance-minimizing estimator subject to the bias constraint. Taken together,

$$\begin{aligned}\inf_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta[r_\theta^D(\hat{\tau}_\pi(r^I))] &\geq \liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] \\ &\geq \mathrm{E}_\eta \min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}) + \mathrm{E}_\eta \bar{\beta}_\theta^2.\end{aligned}$$

Now, by aligned delegation,

$$\min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi(\operatorname{Var}_\theta(\hat{\tau}) + \bar{\beta}_\theta^2) = \min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi r_\theta^D(\hat{\tau}) = \mathrm{E}_\pi r_\theta^D(\hat{\tau}_\pi(r^I))$$

for every $r^I \in \mathcal{R}^*$ for choices from $\mathcal{C}_{\bar{\beta}}$. It follows that for every mechanism there is a set of biases such that the fixed-bias mechanisms has at least weakly better worst-case (over investigator types in $\mathcal{R}^*$) performance. Hence, at an optimal choice of biases β^η given the hyperprior η, the fixed-bias restriction $\mathcal{C}^\eta$ is minimax optimal. Such a minimizer exists because the set of biases is wlog compact (indeed, we can assume $\mathrm{E}_\eta \beta_\theta^2 \leq \mathrm{E}_\eta r_\theta^D(z \mapsto 0) < \infty$) and the expected average risk continuous in the

choice of bias. □

I conjecture that the restriction of the support to priors with full support is not necessary.

B Representation of Unbiased Estimators

As in the main text, for fixed $n \geq 1$ and finite support $\mathcal{Y}$ I consider potential outcomes $\theta = (y(1), y(0)) \in \Theta = (\mathcal{Y}^2)^n$ from which for treatment $d \in \{0,1\}^n$ we observe $y = d \circ y(1) + (\mathbf{1} - d) \circ y(0) \in \mathcal{Y}^n$. (Here, $\circ$ denotes the Hadamard (entry-wise) product.) The estimate of interest is $\tau_\theta = \mathbf{1}'(y(1) - y(0))/n$.

Lemma 1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

1. *For a known treatment probability p, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}' d_{-ij} = n_1 - 1$.

I build up this general representation result in steps from simple estimators with binary outcomes to general estimators with finite support.

B.1 Known treatment probability, binary outcomes

I start with known treatment probability $p = \mathrm{E}_\theta[d_i]$ with d_i iid and binary support.

A natural class of admissible estimators are Bayes estimators, so a tempting starting point for the analysis of optimal unbiased estimators are (limits of) Bayes estimators that minimize average mean-squared error given the data and are also unbiased. However:

Remark B.1. *For $\mathcal{Y} = \{0, 1\}$ and $p = .5$, the only unconstrained Bayes estimator (with respect to average mean-squared error) that is unbiased (conditional on $(y(1), y(0))$) is $\hat{\tau}(y, d) = \frac{1}{n}(2d - \mathbf{1})'(2y - \mathbf{1})$. For $\mathcal{Y} = \{0, 1\}$ and $p \neq .5$, there are no unconstrained Bayes estimators that are also unbiased.*

Sketch of proof. For any prior, the unconstrained Bayes estimator with respect to average mean-squared error is the posterior expectation of τ_θ given the data. Any posterior expectation of τ_θ is bounded between the maximal treatment effect $+1$ and the minimal treatment effect -1. To achieve unbiasedness, any data that is consistent with either of the extremes must therefore yield an estimate of $+1$ or -1, respectively. Iterating this argument, the unique unconstrained Bayes estimator is the one achieved from a prior that puts full probability on $(y_i(1), y_i(0)) \in \{(1, 0), (0, 1)\}$ and zero probability on the configurations $\{(1, 1), (0, 0)\}$. This yields $\mathrm{E}_\theta[y_i(1) - y_i(0)|y_i, d_i] = (2d_i - 1)(2y_i - 1)$, which is unbiased for $p = .5$, but not for $p \neq .5$. □

The remark implies that searching for unbiased estimators among unconstrained Bayes estimators to characterize the class of admissible unbiased estimators is futile, and I instead first characterize unbiased estimators before returning to optimality by solving for constrained Bayes estimators subject to the resulting representation.

Theorem B.1. *For $\mathcal{Y} = \{0, 1\}$, assume that the estimators $\hat{\tau}^A, \hat{\tau}^B$ are unbiased τ_θ (conditional on $\theta = (y(1), y(0))$). Then,*

$$\hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} \phi_i(y_{-i}, d_{-i})$$

for a set of functions $\phi_i : (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R}$.

For $n = 2$, the proof of Theorem B.1 can be made on a two-dimensional lattice folded into a torus. The general proof can similarly be understood as summing over hypercubes on the surface of an n-torus.

Proof. For $\hat{\delta}(y, d) = \hat{\tau}^B(y, d) - \hat{\tau}^A(y, d)$, take $\phi_i(y_{-i}, d_{-i})$ such that

$$\hat{\delta}(y, d) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} \phi_i(y_{-i}, d_{-i}) \tag{9}$$

for all (y, d) with $y'd > 0$ (that is, all those that include some pair $(y_j, d_j) = (1, 1)$). This is always feasible, say by the following inductive construction:

1. Set the $\phi_i(\mathbf{1}_{n-1}, \mathbf{1}_{n-1})$ in any way that has (9) hold for $\hat{\delta}(\mathbf{1}_n, \mathbf{1}_n)$.

2. Assuming that $\phi_i(y_{-i}, d_{-i})$ has been set for all i and (y, d) with $y'd \geq n - k$ such that (9) holds for such (y, d) (as is the case for $k = 0$ by the previous step), consider (y, d) with $y'd = n-(k+1)$. Among the terms $\phi_i(y_{-i}, d_{-i})$ in (9), those with $y'_{-i}d_{-i} = n - (k + 1)$ have already been set by the induction assumption, and it remains to show that we can set conformable terms $\phi_i(y_{-i}, d_{-i})$ for $y'_{-i}d_{-i} = n - (k + 2)$.

   Provided that $k < n - 1$, note that any (y, d) with $y'd = n - (k + 1)$ contains at least one (y_i, d_i) with $y'_i d_i = 1$, $\hat{\delta}(y, d)$ has the term $\phi_i(y_{-i}, d_{-i})$ appear on the right in (9), where thus $y'_{-i}d_{-i} = y'd - 1 = n - (k + 2)$ (so it has not yet been set). But note that this specific $\phi_i(y_{-i}, d_{-i})$ also appears only for that (y, d) among all (y, d) with $y'd = n - (k + 1)$ as necessarily $y'_i d_i = 1$. Hence, we can set all previously undetermined $\phi_i(y_{-i}, d_{-i})$ for all i and $y'd$ with $y'd \geq n - (k + 1)$ in a way that (9) holds for such (y, d).

By induction, we have set all $\phi_i(y_{-i}, d_{-i})$ for any i and $y'd \geq 1$ conformably with (9) for such (y, d). since this includes *all* terms of the form $\phi_i(y_{-i}, d_{-i})$, it remains to show that the unbiasedness assumption implies that (9) extends to (y, d) with $y'd = 0$.

Write $\hat{\delta}^\phi$ for the function defined by (9) for all (y, d). We have thus shown that $\hat{\delta}^\phi(y, d) = \hat{\delta}(y, d)$ for all (y, d) with $y'd > 0$. By assumption, $\mathrm{E}_\theta[\hat{\delta}(y, d)] = 0$ for all $\theta = (y(1), y(0))$, so

$$0 = \mathrm{E}_\theta[\hat{\delta}(y, d)] = \sum_{d \in \{0,1\}^n} \mathrm{P}(d)\, \hat{\delta}(d \circ y(1) + (1 - d) \circ y(0), d).$$

Fixing (y^*, d^*), it follows for any $\tilde{y}$ that

$$\hat{\delta}(y^*, d^*) = -\sum_{d \in \{0,1\}^n \setminus \{d^*\}} \mathrm{P}(d)/\,\mathrm{P}(d^*)\, \hat{\delta}((\mathbb{1}_{d_i = d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d) \tag{10}$$

Since $\hat{\delta}^\phi$ is similarly zero-bias by construction, the same holds for $\hat{\delta}^\phi$. Thus, if for

some (y^*, d^*) $\hat{\delta}$ and $\hat{\delta}^\phi$ agree on

$$\tilde{y}^*(d) = (\mathbb{1}_{d_i = d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d)$$

for some $\tilde{y}$ and all $d \neq d^*$, then $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$.

We are ready to show (9) for all (y^*, d^*), by induction over $\mathbf{1}'d^*$. We let $\tilde{y} = \mathbf{1}$ throughout. At $k = 0$, $d^* = \mathbf{0}$. For any $d \neq d^*$, $\tilde{y}^*(d)'d \geq 1$, so $\hat{\delta}(\tilde{y}^*(d), d) = \hat{\delta}^\phi(\tilde{y}^*(d), d)$. By (10), $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$. Assume now that the claim holds for all (y^*, d^*) with $\mathbf{1}'d^* \leq k$, and consider some (y^*, d^*) with $\mathbf{1}'d^* = k+1$. Then, for any $d \neq d^*$ with $\mathbf{1}'d \leq k$, $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ by the induction assumption. For any $d \neq d^*$ with $\mathbf{1}'d \geq k+1$ there must be at least one dimension i with $d_i = 1, d_i^* = 0$, thus $\tilde{y}^*(d)'d \geq 1$ and $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ follows by construction. We conclude that $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ for all (y^*, d^*). □

Since $\hat{\tau}(y, d) = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)} y_i$ is unbiased for τ_θ, the following characterization is immediate:

Corollary B.1. *For $\mathcal{Y} = \{0, 1\}$, any unbiased estimator $\hat{\tau}$ of τ_θ can be expressed as*

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)}(y_i - \phi_i(y_{-i}, d_{-i})).$$

The following result for the special case $n = 2$ shows that the reduction in degrees of freedom in the estimator implied by unbiasedness is substantial:

Remark B.2. *For $n = 2$, the $\phi_i(y_{-i}, d_{-i})$ are unique up to the one-dimensional equivalence class $\phi_i'(y_{-i}, d_{-i})) = \phi_i'(y_{-i}, d_{-i})) + (-1)^i(2d_{3-i} - 1)\Delta$, so unbiasedness reduces the degrees of freedom from $\hat{\tau} \in \mathbb{R}^{16}$ to $[\phi] \in \mathbb{R}^7$.*

B.2 Fixed treatment group size, binary

Assume now that instead of the treatment probability, the number of treated is fixed at n_1, so that $d \sim \mathcal{U}(\mathcal{D}_{n_1})$ with $\mathcal{D}_{n_1} = \{t \in \{0, 1\}^n; t'n = n_1\}$. Effectively, we assume invariance to permutations in the assignment of treatment, but not more.

The natural, unbiased treatment-control-difference estimator can be written as

$$\hat{\tau}^*(y, d) = \frac{1}{n_1}\sum_{d_i=1} y_i - \frac{1}{n_0}\sum_{d_i=0} y_i = \frac{1}{n_1 n_0}\sum_{d_i=1, d_j=0} (y_i - y_j),$$

of which an unbiased extension is

$$\hat{\tau}^{\phi}(y,d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j - \phi_{ij}(y_{-ij}, d_{-ij}))$$

with $\phi_{ij} = -\phi_{ji}$. I claim that these are also *all* extensions.

Theorem B.2. *Let $\mathcal{Y} = \{0,1\}$. Assume that $\hat{\tau}^A, \hat{\tau}^B$ are unbiased for τ_θ. Then,*

$$\hat{\tau}^B(y,d) - \hat{\tau}^A(y,d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} \phi_{ij}(y_{-ij}, d_{-ij}), \qquad \phi_{ij} = -\phi_{ji}$$

for functions $\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R}$.

Note that we can alternatively write

$$\hat{\tau}^B(y,d) - \hat{\tau}^A(y,d) = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_i - d_j)\, \phi_{ij}(y_{-ij}, d_{-ij}),$$

where we sum over each pair once and ϕ_{ij} is only defined for $j > i$.

We first establish a lemma that adopts the proof strategy from Theorem B.1 to the setting at hand. To this end, for $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ write

$$N(y(1), y(0)) = \{(d \circ y(1) + (1-d) \circ y(0), d); d \in \mathcal{D}_{n_1}\}$$

(the set of observations consistent with $y(1), y(0)$) and let

$$\mathcal{C} = \bigcup_{(y(1),y(0)) \in (\mathcal{Y}^2)^n} N(y(1), y(0)).$$

Let $c : \mathcal{C} \to \mathcal{C}^-$ be the surjective correspondence

$$(y,d) \mapsto \{(ij, (y_{-ij}, d_{-ij})); i < j, d_i \neq d_j\}.$$

Lemma B.1. *If there exists a partition $\mathcal{C} = \bigcup_{t=1}^{T} \mathcal{C}_t$ such that for some T^**

1. *for $\mathcal{C}_t^- = \bigcup_{(y,d) \in \mathcal{C}_t} c(y,d)$ and*

$$\mathcal{D}_t = \mathcal{C}_t^- \setminus \bigcup_{s<t} \mathcal{C}_s^-,$$

there exists injections $b_t : \mathcal{C}_t \to \mathcal{D}_t$ for $t \leq T^*$ and

2. *for all* $t > T^*$ *and* $(y,d) \in \mathcal{C}_t$*, there exists some* $(y(1),y(0)) \in (\mathcal{Y}^2)^n$ *both* $(y,d) \in N(y(1),y(0))$ *and*

$$(N(y(1),y(0)) \setminus \{(y,d)\}) \cap \bigcup_{s \geq t} \mathcal{C}_s = \emptyset$$

then for any $\hat{\delta}$ *that is mean-zero there exist a function* $\phi : \mathcal{C}^- \to \mathbb{R}$ *such that* $\hat{\delta} = \hat{\delta}^\phi$ *with*

$$\hat{\delta}^\phi(y,d) = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_i - d_j)\, \phi_{ij}(y_{-ij}, d_{-ij}).$$

Proof. Given some $\hat{\delta}$, we first construct such a family ϕ with $\hat{\delta}^\phi(y,d) = \hat{\delta}$ for all $(y,d) \in \bigcup_{t \leq T^*} \mathcal{C}_t$, and then establish that this implies $\hat{\delta}^\phi(y,d) = \hat{\delta}$ also for $(y,d) \in \bigcup_{t > T^*} \mathcal{C}_t$.

For the first part, I argue inductively as follows: Take $t \leq T^*$ and assume ϕ has been set on $\bigcup_{s<t} \mathcal{C}_s^-$ such that $\hat{\delta}^\phi = \hat{\delta}$ on $\bigcup_{s<t} \mathcal{C}_s$ (which is given trivially for $t = 1$) then for every $(y,d) \in \mathcal{C}_t$ by the first assumption of the lemma there exists a unique term $\phi_{ij}(y_{-ij}, d_{-ij}) = \phi(b_t(y,d))$ with $b_t(y,d) \in \mathcal{D}_t$ that has not yet been set, so we can set the terms $\phi(\mathcal{D}_t)$ in a way that $\hat{\delta}^\phi = \hat{\delta}$ on $\mathcal{C}_t$ and thus on $\bigcup_{s \leq t} \mathcal{C}_s$. This completes the proof of the first part.

For the second part, note that by assumption $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1),y(0))$, so

$$0 = \mathrm{E}_\theta[\hat{\delta}(y,d)] = \sum_{(y,d) \in N(y(1),y(0))} \hat{\delta}(y,d).$$

Fixing (y^*,d^*) it follows for any $(y(1),y(0))$ with $(y^*,d^*) \in N(y(1),y(0))$ that

$$\hat{\delta}(y^*,d^*) = -\sum_{(y,d) \in N(y(1),y(0)) \setminus \{(y^*,d^*)\}} \hat{\delta}(y,d) \tag{11}$$

Since $\hat{\delta}^\phi$ is similarly zero-bias by construction, the same holds for $\hat{\delta}^\phi$. We are now ready to show that $\hat{\delta}^\phi = \hat{\delta}$ for all $(y,d) \in \mathcal{C}_t$, by induction over t. For some $t > T^*$, assuming $\hat{\delta}^\phi = \hat{\delta}$ holds for all $(y,d) \in \mathcal{C}_s$ with $s < t$ (as is the case for all $s \leq T^*$), take any $(y^*,d^*) \in \mathcal{C}_t$. By the second part of the lemma, (11) and the induction assumption we must have $\hat{\delta}(y^*,d^*) = \hat{\delta}^\phi(y^*,d^*)$. This completes the proof. □

We are ready to prove the main result:

Proof of Theorem B.2. $\hat{\delta}(y,d) = \hat{\tau}^B(y,d) - \hat{\tau}^A(y,d)$ is a unbiased estimator of zero. Define $a, b : \mathcal{C} \to \mathbb{N}_0$ by

$$a(y,d) = y'd, \qquad\qquad b(y,d) = (\mathbf{1}-y)'(\mathbf{1}-d).$$

Note that $a(y,d) + b(y,d) \leq n$.

First, set $T^* = n-1$ and for every $t \leq T$

$$\mathcal{C}_t = \{(y,d) \in \mathcal{C}; \min(a(y,d), b(y,d)) \geq 1, a(y,d) + b(y,d) = n+1-t\}.$$

Then the first assumption of Lemma B.1 is fulfilled, as for every $(y,d) \in \mathcal{C}_t$ there exists some $(ij, (y_{-ij}, d_{-ij}) \in \mathcal{C}_t$ with $y'_{-ij}d_{-ij} + (\mathbf{1}-y_{-ij})'(\mathbf{1}-d_{-ij}) = n-1-t = a(y,d)+b(y,d)-2$, but (y,d) is also the unique element in $\mathcal{C}_t$ covering that element of $\mathcal{D}_t$ under the correspondence c (as indeed necessarily $y_i = d_i, y_j = d_j$, which pins down (y,d) from $(ij,(y_{-ij},d_{-ij}))$.

Second, with $T = n+1$ and

$$\begin{aligned}\mathcal{C}_n &= \{(y,d) \in \mathcal{C}; a(y,d) = 0, b(y,d) \geq 1\},\\ \mathcal{C}_{n+1} &= \{(y,d) \in \mathcal{C}; b(y,d) = 0\},\end{aligned}$$

note that for each $(y^*,d^*) \in \mathcal{C}_n \cup \mathcal{C}_{n+1}$ we have that $(y(1),y(0)) = (y^* \circ d^* + \mathbf{1} \circ (1-d^*), y^* \circ (1-d^*))$ produces

$$N(y(1),y(0)) \cap \{(y,d) \in \mathcal{C}; \min(a(y,d),b(y,d)) = 0\} = \{(y^*,d^*)\}$$

for $(y^*,d^*) \in \mathcal{C}_n$ and

$$N(y(1),y(0)) \cap \{(y,d) \in \mathcal{C}; b(y,d) = 0\} = \{(y^*,d^*)\}$$

for $(y^*,d^*) \in \mathcal{C}_{n+1}$. This verifies the second assumption of Lemma B.1. □

Unbiased estimators (for binary outcomes) are thus fully characterized by leave-two-out adjustments. Note that leave-one-out adjustments as in the case of known treatment probability p would not generally be unbiased.

B.3 Extension to finite support

Take some distribution over the treatment assignment vector $d \in \{0,1\}^n$, data $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ as before where $\mathcal{Y} \subseteq \mathbb{R}$, and $y = d \circ y(1) + (\mathbf{1} - d) \circ y(0)$. Our goal now is to extend a representation for binary outcomes to one for finite (but arbitrarily large) support $\mathcal{Y}$.

Lemma B.2. *Assume that for $\mathcal{Y} = \{0,1\}$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\delta(y,d)] = 0$ for all $\theta = (y(1), y(0))$ permits a representation $\hat{\delta} = \hat{\delta}^\phi$ with*

$$\hat{\delta}^\phi(y,d) = \sum_{\iota \in \mathcal{I}} w_\iota(d_{S_\iota}) \phi_\iota(y_{-S_\iota}, d_{-S_\iota})$$

for fixed $\mathcal{I}, (w_\iota)_{\iota\in\mathcal{I}}, (S_\iota)_{\iota\in\mathcal{I}}$ (where $\mathcal{I}$ finite) and variable $(\phi_\iota)_{\iota\in\mathcal{I}}$ where

$$\phi_\iota : (\mathcal{Y} \times \{0,1\})^{\{1,\dots,n\}\setminus S_\iota} \to \mathbb{R}.$$

Then the representation result extends to any finite $\mathcal{Y} \subseteq \mathbb{R}$ (with the same $\mathcal{I}, (w_\iota)_{\iota\in\mathcal{I}}, (S_\iota)_{\iota\in\mathcal{I}}$).

Proof. Write $\mathcal{Y}_\ell = \{0, 1, \dots, \ell\}$ and define (for $\ell \geq 2, m \geq 0$)

$$\mathcal{Y}_{\ell,m} = \bigtimes_{i=1}^{m} \mathcal{Y}_{2\ell-1} \times \bigtimes_{i=m+1}^{n} \mathcal{Y}_\ell$$

We first establish the following intermediate result by induction over $t = ns + m$ from $t = 0$: For any $(s,m) \in (\mathbb{N}_0 \times \{1,\dots,n\}) \cup \{(0,0)\}$ for $\ell = 2^s + 1$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1), y(0)) \in \mathcal{Y}_{\ell,m}^2$ permits a representation $\hat{\delta} = \hat{\delta}^\phi$ as above with $\phi_\iota : \bigtimes_{i\in\{1,\dots,n\}\setminus S_\iota} (\mathcal{Y}_{\ell,m})_i \to \mathbb{R}$

For $t = 0$, the statement holds by the assumption of the lemma. Assume now that its holds for t with such (s,m) such that $t = ns+m$ and $\ell = 2^s+1$, and consider the $(s^+, m^+) \in \mathbb{N}_0 \times \{1,\dots,n\}$ with $ns^+ + m^+ = t + 1$, and write $\ell^+ = 2^{s^+} + 1$ Fix an estimator $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1), y(0)) \in \mathcal{Y}_{\ell^+,m^+}^2$. Write For $(y,d) \in \mathcal{Y}_{\ell,m} \times \{0,1\}^n$ define $y^+_{m^+} = \ell^+ + y_{m^+} - 1, y^+_{-m^+} = y_{-m^+}$ as well as $y^-_{m^+} = \ell^+, y^-_{-m^+} = y_{-m^+}$ to obtain $y^+, y^- \in \mathcal{Y}_{\ell^+,m^+}$, and define estimators by

$$\hat{\delta}_1(y,d) = \hat{\delta}(y^+,d) - \hat{\delta}(y^-,d) \qquad\qquad \hat{\delta}_2(y,d) = \hat{\delta}(y,d)$$

where thus $\hat{\delta}_2$ is merely a restriction of $\hat{\delta}$ to $\mathcal{Y}_{\ell,m} \times \{0,1\}^n$. For $(y,d) \in \mathcal{Y}_{\ell^+,m^+} \times$

$\{0,1\}^n$ define $\bar{y}_{m^+} = \min(y_{m^+}, \ell^+), \bar{y}_{-m^+} = y_{-m^+}$ to obtain a $\bar{y} \in \mathcal{Y}^2_{\ell,m}$ for which

$$\begin{aligned}\hat{\delta}(y,d) &= \hat{\delta}(y,d) - \hat{\delta}(\bar{y}_{m^+},d) + \hat{\delta}(\bar{y}_{m^+},d)\\ &= \hat{\delta}_1(y-\bar{y}_{m^+},d) + \hat{\delta}_2(\bar{y}_{m^+},d).\end{aligned}$$

$\hat{\delta}_2$ is unbiased (for $\mathcal{Y}_{\ell,m}$) by construction. Note that

$$\mathrm{E}_\theta[\hat{\delta}_1(y,d)] = \mathrm{E}_\theta[\hat{\delta}(y^+,d)] - \mathrm{E}_\theta[\hat{\delta}(y^-,d)] = 0$$

for any $y(1), y(0) \in \mathcal{Y}_{\ell,m}$, as they generate $y^+(1), y^+(0) \in \mathcal{Y}_{\ell^+,m^+}$ for which $\hat{\delta}$ is unbiased by assumption, so $\hat{\delta}_1$ is likewise unbiased (for $y(1), y(0) \in \mathcal{Y}_{\ell,m}$). By the induction assumption, there are thus ϕ^1, ϕ^2 with

$$\hat{\delta}(y,d) = \sum_{\iota \in \mathcal{I}} w_\iota(d_{S_\iota})(\phi^1_\iota((y-\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota}) + \phi^2_\iota((\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota})$$

for any $(y,d) \in \mathcal{Y}_{\ell^+,m^+} \times \{0,1\}^n$. For

$$\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \phi^1_\iota(y_{-S_\iota} - (\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota}) + \phi^2_\iota((\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota})$$

we therefore have $\hat{\delta} = \hat{\delta}^\phi$. This concludes the induction step and thus the proof of the intermediate result.

Setting $m = n$, it is immediate that the statement of the lemma holds for all $\mathcal{Y} = \mathcal{Y}_{2^s+1}$. Since it will always hold for subsets, it holds for all $\mathcal{Y} = \mathcal{Y}_\ell$. Now take arbitrary $\mathcal{Y} = \{z_1, \ldots, z_\ell\}$, and define for $(y,d) \in (\mathcal{Y}_\ell \times \{0,1\})^n$

$$\tilde{\delta}(y,d) = \hat{\delta}(z_y, d)$$

where $(z_y)_i = z_{y_i} \in \mathcal{Y}$. By the intermediate result there is some $\tilde{\phi}$ such that $\tilde{\delta} = \hat{\delta}^{\tilde{\phi}}$. Setting $\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \tilde{\phi}(\tilde{y}_{-S_\iota}, d_{-S_\iota})$ with $\tilde{y}$ such that $z_{\tilde{y}} = y$ yields $\hat{\delta}(y,d) = \hat{\delta}^\phi(y,d)$. □

We are now ready to proof the representation result in the main paper.

Proof of Lemma 1. The representation for general finite support follows from Lemma B.2 applied to the binary representation results in Theorem B.1 and Theorem B.2, respectively. □

C Characterization of Optimal Unbiased Estimators

When is an estimator not just unbiased, but has also low average mean-squared error? I start with the representation

$$\hat{\tau}^{\phi}(y,d) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \phi_i(y_{-i}, d_{-i}))$$

for known treatment probability p and consider the error

$$\begin{aligned}\Delta_{\theta}^{\phi}(y,d) &= \hat{\tau}^{\phi}(y,d) - \tau_{\theta} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i - p}{p(1-p)}(y_i - \phi_i(y_{-i}, d_{-i})) - (y(1)_i - y(0)_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(\bar{y}_i - \phi_i(y_{-i}, d_{-i}))\end{aligned}$$

for the adjustment oracle $\bar{y}_i = (1-p)y(1)_i + py(0)_i$, which would be the loss-minimizing choice for $\phi_i(y_{-i}, d_{-i})$.

Proposition C.1. *For some prior π over $\theta = (y(1), y(0))$, any ϕ_{π}^{*} with*

$$\phi_{\pi}^{*}(y_{-i}, d_{-i}) = \mathrm{E}_{\pi}\left[\bar{y}_i | y_{-i}, d_{-i}\right]$$

is a (global) minimizer of average loss $\mathrm{E}_{\pi} L_{\theta}(\phi)$, where $L_{\theta}(\phi) = \mathrm{E}_{\theta}(\Delta_{\theta}^{\phi}(y,d))^2$.

Proof. The restriction that adjustments $\phi_i(y_{-i}, d_{-i})$ are functions only of y_{-i}, d_{-i} (and of π) requires some care, as each such adjustments appears given multiple draws of (y,d). Write

$$M_i(y_{-i}^{*}, d_{-i}^{*}) = \{(y,d) \in (\mathcal{Y} \times \{0,1\})^n; (y_{-i}, d_{-i}) = (y_{-i}^{*}, d_{-i}^{*})\}$$

for the (y,d) for which $\hat{\tau}^{\phi}(y,d)$ (and thus $\Delta_{\theta}^{\phi}(y,d)$) includes the term $\phi_i(y_{-i}^{*}, d_{-i}^{*})$. Then,

$$\begin{aligned}\frac{\partial \mathrm{E}_{\pi} L_{\theta}(\phi)}{\partial \phi_i(y_{-i}^{*}, d_{-i}^{*})} &= \frac{\partial \mathrm{E}_{\pi}\left[\mathbb{1}_{(y,d)\in M(y_{-i}^{*}, d_{-i}^{*})}(\Delta_{\theta}^{\phi}(y,d))^2\right]}{\partial \phi_i(y_{-i}^{*}, d_{-i}^{*})} \\ &= \mathrm{E}_{\pi}\left[\mathbb{1}_{(y,d)\in M(y_{-i}^{*}, d_{-i}^{*})} \frac{\partial (\Delta_{\theta}^{\phi}(y,d))^2}{\partial \phi_i(y_{-i}^{*}, d_{-i}^{*})}\right],\end{aligned}$$

where we note that we can exchange differentiation and integration because all summands are bounded. I omit writing E_θ explicitly inside E_π and consider the joint distribution of θ and z. Here, for all $(y,d) \in M(y^*_{-i}, d^*_{-i})$,

$$\begin{aligned}
&\frac{\partial(\Delta^\phi_\theta(y,d))^2}{\partial\phi_i(y^*_{-i},d^*_{-i})} = -\frac{2}{n}\frac{d_i-p}{p(1-p)}\Delta^\phi_\theta(y,d) \\
&= -\frac{2}{n^2}\left(\frac{(d_i-p)^2}{(p(1-p))^2}(\bar{y}_i - \phi_i(y^*_{-i},d^*_{-i})) + \sum_{j\neq i}\frac{(d_i-p)(d^*_j-p)}{(p(1-p))^2}(\bar{y}_j - \phi_j(y_{-j},d_{-j}))\right).
\end{aligned}$$

The first-order condition $\frac{\partial \mathrm{E}_\pi L_\theta(\phi)}{\partial\phi_i(y^*_{-i},d^*_{-i})} = 0$ is therefore

$$\begin{aligned}
&\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i},d^*_{-i})}(d_i-p)^2(\phi_i(y^*_{-i},d^*_{-i}) - \bar{y}_i)\right] \\
&= -\sum_{j\neq i}(d^*_j - p)\,\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i},d^*_{-i})}(d_i-p)(\phi_j(y_{-j},d_{-j}) - \bar{y}_j)\right].
\end{aligned}$$

The condition is trivially fulfilled for $\mathrm{P}_\pi((y,d) \in M(y^*_{-i},d^*_{-i})) = 0$. Otherwise, equivalently

$$\begin{aligned}
&\overbrace{\mathrm{E}[(d_i-p)^2]\phi_i(y^*_{-i},d^*_{-i}) - \mathrm{E}_\pi[(d_i-p)^2\bar{y}_i|(y_{-i},d_{-i}) = (y^*_{-i},d^*_{-i})]}^{=p(1-p)(\phi_i(y^*_{-i},d^*_{-i}) - \mathrm{E}_\pi[\bar{y}_i|(y_{-i},d_{-i})=(y^*_{-i},d^*_{-i})])} \\
&= -\sum_{j\neq i}(2d^*_j - 1)\,\mathrm{E}_\pi\left[(d_i-p)\phi_j(y_{-j},d_{-j})|(y_{-i},d_{-i}) = (y^*_{-i},d^*_{-i})\right]
\end{aligned}$$

Note that this system of first-order conditions will generally have many solutions, as the ϕ-representation of $\hat{\tau}^\phi$ is not generally unique. I now show that the specific choice

$$\phi_i(y^*_{-i},d^*_{-i}) = \mathrm{E}_\pi[\bar{y}_i|(y_{-i},d_{-i}) = (y^*_{-i},d^*_{-i})]$$

(for $\mathrm{E}_\pi \mathrm{P}_d((y,d) \in M(y^*_{-i},d^*_{-i})) > 0$, otherwise, say, zero) is a (global) posterior-loss

minimizer. To that end, note that for $i \neq j$

$$\begin{aligned}
&\mathrm{E}_\pi\left[(d_i - p)\,\mathrm{E}_\pi[\bar{y}_j|y_{-j}, d_{-j}]|y_{-i}, d_{-i}\right] \\
&= \mathrm{E}_\pi\left[(d_i - p)\,\mathrm{E}_\pi[\bar{y}_j|y_i, d_i, y_{-ij}, d_{-ij}]|y_j, d_j, y_{-ij}, d_{-ij}\right] \\
&= \mathrm{E}_\pi\left[\mathrm{E}_\pi\left[(d_i - p)\,\mathrm{E}_\pi[\bar{y}_j|y_i, d_i, y_{-ij}, d_{-ij}]|d_i, y_{-ij}, d_{-ij}\right]|y_{-ij}, d_{-ij}\right] \\
&= \mathrm{E}_\pi\left[(d_i - p)\,\mathrm{E}_\pi[\bar{y}_j|d_i, y_{-ij}, d_{-ij}]|y_{-ij}, d_{-ij}\right] \\
&= \mathrm{E}_\pi\left[(d_i - p)\,\mathrm{E}_\pi[\bar{y}_j|y_{-ij}, d_{-ij}]|y_{-ij}, d_{-ij}\right] = 0.
\end{aligned}$$

The first-order condition follows. Also

$$\begin{aligned}
&\frac{\partial^2\,\mathrm{E}_\pi L_\theta(\phi)}{\partial\phi_i(y^A_{-i}, d^A_{-i})\partial\phi_j(y^B_{-j}, d^B_{-j})} \\
&= \frac{1}{(p(1-p)n)^2}\,\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^A_{-i},d^A_{-i})\cap M(y^B_{-j},d^B_{-j})}(d^B_i - p)(d^A_j - p)\right] \\
&= \begin{cases}
\frac{1}{p(1-p)n^2}\,\mathrm{P}_\pi((y_{-i}, d_{-i}) = (y^A_{-i}, d^A_{-i})), & (i, y^A_{-i}, d^A_{-i}) = (j, y^B_{-j}, d^B_{-j}) \\
\frac{(d^*_i - p)(d^*_j - p)}{(p(1-p)n)^2}\,\mathrm{P}_\pi(y^*, d^*), & i \neq j, (y^{A/B}_{-i/j}, x^{A/B}_{-i/j}) = (y^*_{-i/j}, d^*_{-i/j}) \\
0, & \text{otherwise.}
\end{cases}
\end{aligned}$$

Note that $\frac{\partial^2\,\mathrm{E}_\pi L_\theta(\phi)}{\partial\phi_i(y^A_{-i},d^A_{-i})\partial\phi_j(y^B_{-j},d^B_{-j})}$ is two times the variance-covariance matrix of the (mean-zero) random variables $\mathbb{1}_{(y,d)\in M(y^*_{-i},d^*_{-i})}\frac{d_i - p}{p(1-p)n}$, and therefore everywhere positive semi-definite. It follows that the first-order conditions locate a (global) minimum. □

The proposition directly yields the first part of the general characterization result in the main paper. The second part follows analogously with oracle adjustments

$$\Delta\bar{y}_{ij} = \underbrace{\left(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\right)}_{\bar{\bar{y}})_i} - \left(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\right).$$

Theorem 2 (Solution of the investigator)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i|z_{-i}]).$$

2. For a fixed number n_1 of treated units,

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta \bar{y}_{ij} | z_{-ij}]).$$

Proof. The first part is immediate from Proposition C.1. For the second part, we can wlog consider adjustments

$$\phi_{i;j}(y_{-ij}, d_{-ij}) \tag{12}$$

for which we set $\phi_{ij}(y_{-ij}, d_{-ij}) = \phi_{i;j}(y_{-ij}, d_{-ij}) - \phi_{j;i}(y_{-ij}, d_{-ij})$ to find

$$\begin{aligned}
\Delta_\theta^\phi(y, d) &= \hat{\tau}^\phi(y, d) - \tau_\theta \\
&= \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j) \left((\bar{\bar{y}}_i - \phi_{i;j}(y_{-ij}, d_{-ij})) - (\bar{\bar{y}}_j - \phi_{j;i}(y_{-ij}, d_{-ij})) \right) \\
&= \frac{1}{n_1 n_0} \sum_{i,j} (d_i - d_j)(\bar{\bar{y}}_i - \phi_{i;j}(y_{-ij}, d_{-ij})).
\end{aligned}$$

As in the proof of Proposition C.1, we can then verify that the choice

$$\phi_{i;j}(y_{-ij}, d_{-ij}) = \mathrm{E}_\pi[\bar{\bar{y}}_i | y_{-ij}, d_{-ij}]$$

fulfils the associated first-order condition. □

D OLS is Biased

Consider a sample of n units (y_i, d_i, x_i), where $d_i \in \{0, 1\}$ are iid given $x_1, \ldots, x_n$ with $\mathrm{P}(d_i = 1) = p \in (0, 1)$.

D.1 Conditional on covariates

Conditional on covariates $x_i = \mathbb{1}_{i=1}$ and for $y_i = x_i d_i$, the sample-averge treatment effect is $\tau = 1/n$ (one for the first unit, zero for all other units). The coefficient $\hat{\tau}^{\mathrm{OLS}}$ on d in a linear regression of y on d and x (with intercept) has expectation $\mathrm{E}[\hat{\tau}^{\mathrm{OLS}} | n_1] = 0$ conditional on any number $1 < n_1 < n - 1$ of treated units. Indeed, x perfectly explains y, so the coefficient on d will always be zero (by Frisch-Waugh or otherwise).

D.2 Over the sampling distribution

Assume that $x_i \in \mathbb{R}^{k_n+1}$ with $\mathrm{P}(x_{i0}) = q \in (0,1)$ and

$$x_{i1}, \dots, x_{ik} | x_{i0} \overset{\text{iid}}{\sim} (1 - x_{i0}) \cdot \mathcal{N}(0,1)$$

(that is, $x_{ij} = 0$ for all $j > 0$ if $x_{i0} = 1$), x_i iid across units. (Alternatively, any non-degenerate distribution will do.) Let $y_i = x_{i0} d_i$. The average treatment effect of d_i on y_i is

$$\tau^{\text{pop}} = \mathrm{E}[y_i | d_i = 1] - \mathrm{E}[y_i | d_i = 0] = q.$$

Let $\hat{\tau}^{\text{OLS}}$ be the coefficient on d in a linear regression of y on d and x (with intercept). For $k_n/n \to \alpha \in (0, 1-q)$ as $n \to \infty$ we also find

$$\hat{\tau}^{\text{OLS}} \xrightarrow{\text{P}} \frac{q}{1-\alpha}.$$

Indeed, writing A_x for the annihilator matrix with respect to x and the intercept, by Frisch-Waugh $\hat{\tau}^{\text{OLS}} = \frac{d' A_x y}{d' A_x d}$ with

$$\begin{aligned}
\mathrm{E}[d' A_x y | x] &= p(1-p)(n_{x=1} - 1), \\
\mathrm{E}[d' A_x d | x] &= p(1-p) \operatorname{trace}(A_x) = p(1-p)(n - k_n - 1).
\end{aligned}$$

By the law of large numbers (where variances are suitably bounded),

$$\begin{aligned}
\frac{d' A_x y}{n} &\xrightarrow{\text{P}} p(1-p) \, \mathrm{E}[n_{x=1}/n] = p(1-p)q, \\
\frac{d' A_x y}{n} &\xrightarrow{\text{P}} p(1-p)(1-\alpha).
\end{aligned}$$

E Asymptotic Inference

In this section, I derive asymptotically valid inference of the average treatment effect. These results deviate from the approach in the main paper in two notable, related ways. First, I assume that potential outcomes and controls themselves are sampled iid from a population distribution, and inference will not condition on their realizations. Second, in order to obtain valid inference, I take large-sample approximations. The estimator of interest is still unbiased in finite samples for the sample-average

treatment effect. But for efficiency and inference I focus on the estimation of the population-average treatment effect in large samples.

Building up to a characterization of the variance of the treatment-effect estimator in terms of out-of-sample prediction quality, I first state an auxiliary remark that will simplify the proof of the main result.

Remark E.1 (K-fold variance bound)**.** *Consider n square-integrable, mean-zero random variables $a_1, \dots, a_n$ and a partition $\bigcup_{k=1}^{K} \mathcal{I}_k = \{1, \dots, n\}$ such that, for all k, $\mathrm{E}[a_i a_j] = 0$ for all $i, j \in \mathcal{I}_k$. Then,*

$$\mathrm{Var}\left(\textstyle\sum_{i=1}^{n} a_i\right) \leq K \displaystyle\sum_{i=1}^{n} \mathrm{Var}(a_i).$$

Proof. By Cauchy-Schwarz, applied once per row, we find that

$$\begin{aligned}\mathrm{Var}\left(\textstyle\sum_{i=1}^{n} a_i\right) = \mathrm{Var}\left(\textstyle\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} a_i\right) &\leq \left(\textstyle\sum_{k=1}^{K} \sqrt{\mathrm{Var}\left(\sum_{i \in \mathcal{I}_k} a_i\right)}\right)^2 \\ &\leq K \sum_{k=1}^{K} \mathrm{Var}\left(\textstyle\sum_{i \in \mathcal{I}_k} a_i\right) = K \displaystyle\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \mathrm{Var}(a_i),\end{aligned}$$

where the last equality follows because increments are uncorrelated within folds. □

I assume that potential outcomes and control variables are drawn iid from a population distribution

$$(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P},$$

treatment is assigned according to a known treatment probability $\mathrm{P}(d_i = 1) = p \in (0, 1)$, and data (y_i, d_i, x_i) obtained from $y_i = y_i(d_i)$.

In this section, I focus on K-fold estimators similar to those in Remark 5.1. Specifically, I assume that a sample of size n is divided into K equally-sized folds

$$\bigcup_{k=1}^{K} \mathcal{I}_k = \{1, \dots, n\}$$

(so I implicitly assume that K divides n). In this setting, I consider the asymptotic

distribution of the estimator

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} (y_i - \hat{f}_k(x_i))$$

of the population-average treatment effect $\tau = \mathrm{E}[y(1) - y(0)]$, where each $\hat{f}_k : \mathcal{X} \to \mathbb{R}$ is fitted only on folds other than $\mathcal{I}_k$. My first result characterizes the asymptotic distribution of $\hat{\tau}$. Throughout, I use indices i and k outside sums for a representative draw from the respective distribution.

Theorem E.1 (Asymptotic distribution of K-fold estimator)**.** *Assume that*

1. $\mathrm{E}[\mathrm{Var}(\hat{f}_k(x_i)|x_i)] \to 0$ *as* $n \to \infty$,
2. $\mathrm{E}\left[\left(\frac{1-p}{p}\right)^{2d_i - 1} (y_i - \hat{f}_k(x_i))^2\right] \to L$ *(where* $i \in \mathcal{I}_k$*), and*
3. $\mathrm{E}[(\hat{f}_k(x_i) - y_i)^{2+\delta}] < C < \infty$ *for some* $\delta, C > 0$.

Then,

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, s^2), \qquad s^2 = \frac{L}{p(1-p)} - \tau^2.$$

Note that the distribution of prediction functions $\hat{f}_k$ will depend on the sample size of the training sample, and thus on n. Furthermore, the result can be extended to the case where the population distribution itself depends on n. While I assume that K is fixed here, the conclusion also holds with K growing provided that $K \, \mathrm{E}[\mathrm{Var}(\hat{f}_k(x_i)|x_i)] \to 0$.

The first condition expresses that the prediction variance vanishes and predictions stabilize in large samples. The second condition defines the asymptotic prediction loss of the algorithm. The third condition is a regularity assumption that will ensure asymptotic convergence. When this condition holds, I do not require the assumption of bounded support of potential outcomes from the main paper. Importantly, I do not assume that the prediction functions approximate the best prediction of y given x or are risk-consistent, only that their variance vanishes.

Proof of Theorem E.1. Write $t_i = \frac{d_i - p}{p(1-p)}$. I decompose

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau) &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (t_i(y_i - \underbrace{\mathrm{E}[\hat{f}_k(x_i)|x_i]}_{=g_n(x_i)}) + t_i(\mathrm{E}[\hat{f}_k(x_i)|x_i] - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} t_i(\hat{f}_k(x_i) - g_n(x_i)).
\end{aligned}$$

For the first part, note that $\mathrm{E}[(t_i(y_i - g_n(x_i)) - \tau)^{2+\delta}]$ is bounded, uniformly in n. Its expectation is zero and its variance is

$$\begin{aligned}
s_n^2 &= \mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \right) = \mathrm{Var}\left(t_i(y_i - g_n(x_i))\right) \\
&= \mathrm{E}\big[\underbrace{t_i^2}_{=\left(\frac{d_i - p}{p(1-p)}\right)^2 = \frac{1}{p(1-p)}\left(\frac{1-p}{p}\right)^{2d_i - 1}} (y_i - g_n(x_i))^2 \big] - (\underbrace{\mathrm{E}[t_i(y_i - g_n(x_i))]}_{=\tau})^2 \\
&= \frac{\mathrm{E}\left[\left(\frac{1-p}{p}\right)^{2d_i - 1} (y_i - g_n(x_i))^2 \right]}{p(1-p)} - \tau^2.
\end{aligned}$$

Hence, by the Lyapunov CLT for triangular arrays,

$$\frac{1}{\sqrt{n s_n^2}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \xrightarrow{d} \mathcal{N}(0, 1).$$

Combining the first two assumptions,

$$\mathrm{E}\left[\left(\frac{1-p}{p} \right)^{2d_i - 1} (y_i - g_n(x_i))^2 \right] \to L,$$

so we obtain that $s_n^2 \to s^2 = \frac{L}{p(1-p)} - \tau^2$ and thus

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \xrightarrow{d} \mathcal{N}(0, s^2).$$

For the second part, by Remark E.1,

$$\begin{aligned}
&\operatorname{Var}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} t_i(\hat{f}_k(x_i) - g_n(x_i))\right) \\
&\leq \frac{K}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \operatorname{Var}\left(t_i(\hat{f}_k(x_i) - g_n(x_i))\right) = K\,\mathrm{E}\left[t_i^2(\hat{f}_k(x_i) - g_n(x_i))^2\right] \\
&= K\,\mathrm{E}\left[\left(\frac{d_i - p}{p(1-p)}\right)^2\right]\mathrm{E}\left[(\hat{f}_k(x_i) - g_n(x_i))^2\right] \\
&= \frac{K}{p(1-p)}\,\mathrm{E}\left[(\hat{f}_k(x_i) - \mathrm{E}[\hat{f}_k(x_i)|x_i])^2\right] = \frac{K}{p(1-p)}\,\mathrm{E}\left[\operatorname{Var}(\hat{f}_k(x_i)|x_i)\right] \longrightarrow 0
\end{aligned}$$

as $n \to \infty$. In particular,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} t_i(\hat{f}_k(x_i) - g_n(x_i)) \xrightarrow{\mathrm{P}} 0.$$

The claim of the theorem follows. □

The asymptotic variance is a function of the expected prediction loss and the treatment effect, and can be estimated consistently from the sample analogs.

Remark E.2 (Asymptotically valid variance estimate)**.** *Under the assumptions of Theorem E.1, the asymptotic variance of $\hat{\tau}$ can be estimated consistently by*

$$\hat{s}^2 = \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left(\frac{d_i - p}{p(1-p)}(y_i - \hat{f}_k(x_i)) - \hat{\tau}\right)^2.$$

As a consequence, we can construct asymptotically valid standard errors and Normal-theory confidence intervals from $\hat{s}^2$. To be more precise, $\frac{\hat{s}}{\sqrt{n}}$ is a valid standard error for $\hat{\tau}$, and

$$[\hat{\tau} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}, \hat{\tau} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}]$$

a $1-\alpha$ confidence interval for τ (where $z_{1-\alpha/2}$ is the $1-\alpha/2$-quantile of the standard Normal distribution).

The asymptotic results extend to the case of fixed n_1 (by setting $p = n_1/n$, provided that $\mathrm{E}[\hat{f}_k(x_i)] \to \mathrm{E}[\bar{y}_i]$), exact cross-fitting as in Remark 5.1 with balanced

folds, and folds that are only approximately of the same size or only approximately balanced.

Now that we have established asymptotically valid inference, I am ready to return to preference alignment.

Remark E.3 (Alignment over precision)**.** *Assume the investigator chooses among unbiased estimators, that is, by Lemma 1 among regression adjustments. Assume further that she constructs regression adjustments in a K-fold procedure with (a sequence of) prediction functions that fulfill the regularity assumptions for asymptotically valid inference in Theorem E.1. Then, if the investigator wants to obtain small standard errors or tight confidence intervals, her choices are aligned with the designer's preference for low mean-squared error $\mathrm{E}[(\hat{\tau} - \tau)^2]$ among these unbiased estimators.*

Proof. The asymptotic distribution of $\hat{\tau}$ as well as the probability limit of $\hat{s}^2$ only depend on the asymptotic loss L, the treatment probability p, and the treatment effect τ. The investigator through her choice of adjustments can only control L, and for these preferences chooses a sequence of prediction functions that minimizes asymptotic prediction loss. This is also the variance-minimizing choice the designer prefers. (Since L is non-random, the specific utility function over the size of standard errors or confidence intervals does not matter here.) □

Note that unbiasedness is crucial to reduce the degrees of freedom over the asymptotic distribution to the variance, with respect to which designer and investigator are aligned. Conversely, designer and investigator may have different preferences over the bias-variance trade-off, so allowing for (asymptotic) bias would break alignment even when the estimator is asymptotically Normal.

By the same argument as in the proof of Remark E.3, choices are also aligned over the power of a test against some null hypothesis. Since the investigator cannot move the expectation of the estimator, the best she can do is to pick a sequence of prediction functions for which the asymptotic loss L is minimal.

Remark E.4 (Alignment over power)**.** *Consider a sequence of population distributions with $\tau_n = \tau_0 + \frac{\delta}{\sqrt{n}}$. Assume that the investigator constructs a one- or two-sided test against the null hypothesis $\tau = \tau_0$ by comparing the test statistic $\frac{\sqrt{n}(\hat{\tau}-\tau_0)}{\hat{s}}$ to the standard Normal distribution, and that the investigator's (sequence of) prediction functions fulfill the regularity assumptions in Theorem E.1. If the investigator has a*

preference for rejecting $\tau = \tau_0$, then her choices are aligned with the designer's goal of minimizing $\mathrm{E}[(\hat{\tau} - \tau)^2]$.

Based on the asymptotic approximation from Theorem E.1, I am now ready to prove the result from the main paper that distribution to two researchers attains asymptotic efficiency.

Remark 6.3 (Semi-parametric efficiency)**.** *If researchers use prediction algorithms* $(A_n : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}, z \mapsto \hat{f}_n)_{n=1}^{\infty}$ *with*

$$\mathrm{E}[(\hat{f}_n(x_i) - \mathrm{E}[\bar{y}_i|x_i])^2] \to 0$$

as $n \to \infty$, then delegation to two researchers with risk functions in $\mathcal{R}^*$ (who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased estimation of τ_θ, and large-sample semi-parametric efficient estimation of τ for the semi-parametric efficiency bound of Hahn (1998).

Proof of Remark 6.3. Similar to the proof of Theorem E.1, again setting $t_i = \frac{d_i - p}{p(1-p)}$, I decomposem, with $K = 2$,

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau) &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (t_i(y_i - \mathrm{E}[\bar{y}_i|x_i]) + t_i(\mathrm{E}[\bar{y}_i|x_i] - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - \mathrm{E}[\bar{y}_i|x_i])) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} t_i(\hat{f}_k(x_i) - \mathrm{E}[\bar{y}_i|x_i])).
\end{aligned}$$

The latter part converges to zero in probability by Remark E.1 as in the proof of Theorem E.1. Since the support of potential outcomes is bounded, the first part converges by the standard CLT to a mean-zero Normal distribution with asymptotic variance

$$\mathrm{Var}(t_i(y_i - \mathrm{E}[\bar{y}_i|x_i])) = \frac{\mathrm{E}\,\mathrm{Var}(y_i(1)|x_i)}{p} + \frac{\mathrm{E}\,\mathrm{Var}(y_i(0)|x_i)}{1-p} + \mathrm{Var}(\mathrm{E}[y_i(1) - y_i(0)|x_i]),$$

which is the efficiency bound of Hahn (1998). □