Toward Trustworthy Causal Inference

by

Yaxuan Huang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Distinguished Professor Bin Yu, Chair
Associate Professor Sam Pimentel, Co-chair
Professor Avi Feller
Assistant Professor Srigokel Upadhyayula

Spring 2026

Toward Trustworthy Causal Inference

Abstract

Toward Trustworthy Causal Inference

by

Yaxuan Huang

Doctor of Philosophy in Statistics

University of California, Berkeley

Distinguished Professor Bin Yu, Chair

Associate Professor Sam Pimentel, Co-chair

This dissertation studies methodological challenges in modern causal inference with a particular focus on the trustworthiness and robustness of estimation methods. While recent advances in machine learning and statistical methodology have vastly expanded the scope and toolkit of causal analysis, many questions remain unresolved. In practice, researchers often make causal conclusions despite unobserved counterfactuals, possible unmeasured confounding, as well as covariate imbalance in observational studies. First, unlike supervised prediction problems, since causal estimands such as treatment effects are never directly observed, model choice and uncertainty quantification become substantially more difficult in the absence of ground truth against which competing methods can be compared. Second, the assumption of no unobserved confounding is implausible in practice and existing sensitivity analysis methods seek to quantify how violations of this assumption could affect causal conclusions. However, these methods are typically designed for specific estimation strategies, motivating the need for a more general framework. Third, existing randomization inference methods create a mismatch between the design stage and the inference stage of matched observational studies, and sensitivity analyses for such studies are often inadequate. In the face of these problems, the three chapters of this dissertation develop methodological tools addressing these difficulties, with the common goal of improving the stability, robustness, and trustworthiness of causal inference in modern data settings.

Chapter 1 develops a framework for stable model screening and uncertainty quantification for conditional average treatment effect (CATE) estimation - PCS-CATE. This is a joint work with Sizhu Lu and Bin Yu. The chapter adapts the Predictability–Computability–Stability (PCS) framework for veridical data science of Yu and Kumbier [99] to causal inference settings. Because standard prediction-based validation is unavailable in causal problems, we adopt subgroup calibration metrics and a multi-metric screening procedure for evaluat-

ing CATE estimators. We further introduce a bootstrap-based uncertainty quantification framework that propagates both algorithmic and data uncertainty while calibrating interval estimates through subgroup coverage criteria. The framework is extended from randomized experiments to observational studies through additional propensity-score model diagnostics and screening procedures. Extensive simulations on synthetic and semi-synthetic data demonstrate that the proposed PCS procedure achieves more stable model screening and reliable interval coverage with mild width inflation relative to existing approaches.

Chapter 2 studies the connection between sensitivity analyses of regression and weighting estimators, and formulates a unified framework of examining the sensitivity to unobserved confounding based on the perturbation of implied weights. This chapter is a joint work with Sam Pimentel and Melody Huang. Existing sensitivity analysis methods are often tailored to specific estimation strategies, such as partial R^2 for linear regressions, or variance-based sensitivity model for IPW estimators. With the implied weight form for regression estimators, we constrain the L_2 distance between implied weights based on observable covariates and oracle weight vector and characterize the resulting worst-case bias. We introduce two complementary formulations, the *restricted* model, which leads to interpretable sensitivity parameters and sharp bias bounds within this structured class, and the *unrestricted* model, which provides bounds that are robust to model misspecification. The framework naturally extends to interacted regressions and other estimators that admit implied weights. Our results provide a unified perspective on sensitivity analysis across regression and weighting estimators and offer a tractable approach to sensitivity analysis in settings with heterogeneous treatment effects or more flexible estimators.

Chapter 3 studies robust design and inference through covariate-adaptive randomization in matched observational studies. This is a published work with Sam Pimentel [65]. This chapter develops randomization-based inference procedures that account for covariate imbalance and propensity score estimation uncertainty in matched designs. The chapter proposes covariate-adaptive randomization inference methods that adjust for residual covariate discrepancies after matching, while also incorporating sensitivity analysis for unobserved confounding. The resulting procedures improve finite-sample validity and robustness relative to standard randomization inference approaches. Through theoretical results, simulations, and empirical case studies, the chapter demonstrates how principled design-based adjustments can strengthen causal conclusions in observational studies.

Taken together, the three chapters contribute to trustworthy causal inference in modern data settings. Across model screening, sensitivity analysis, and inference design, this dissertation emphasizes reality check, stability, and transparent uncertainty quantification as fundamental principles for trustworthy causal conclusions in the presence of unobserved variables, model uncertainty and residual covariate imbalance.

Acknowledgments

I feel incredibly fortunate to have spent my PhD journey in the Department of Statistics at University of California, Berkeley, surrounded by brilliant, supportive, diverse, and collaborative people. This five-year adventure has profoundly shaped me, both as a researcher and as a person.

First and foremost, I would like to thank my incredible advisors, Bin Yu and Sam Pimentel. It has been both an honor and a privilege to be co-advised by them.

My very first conversations with Professor Yu happened before I even joined Berkeley. I was immediately drawn to her unique style: sharp, insightful, straightforward, and encouraging. That first impression led to five years of mentorship and collaboration that profoundly shaped my research taste and way of doing research and thinking. I have always been amazed by the breadth of her knowledge across so many different fields, and even more by her ability to draw connections between ideas and areas. She constantly brings endless new ideas and unique perspectives, shaping the group's research into something both diverse and deeply interconnected. On a personal level, she has also been a role model for me. Her persistence and passion for doing the right thing, conducting meaningful research, and caring about real-world impact have deeply influenced me. She has given me invaluable advice not only on research, but also on how to live and think as a person. Although I still feel far from meeting her standards, I will continue striving toward them. I am especially grateful for the trust and freedom she gave me, allowing me to explore my own interests, enter unfamiliar areas, and grow through curiosity and independence.

I am also deeply grateful to Professor Sam Pimentel, who has always been encouraging, thoughtful, and supportive throughout my PhD journey. His intellectual rigor in causal inference, especially in observational studies, has deeply shaped my statistical thinking. Sam is extraordinarily detail-oriented, structured, and insightful. He introduced me to a more rigorous statistical world through constructive feedback, thoughtful questions, and constant encouragement. When I had a clear direction, he provided support and guidance; when I felt lost, he encouraged me and helped me regain confidence and clarity. No matter how my projects were going, he was always supportive, which meant especially much during the difficult periods of my PhD. Beyond research, I also learned from him how to be organized, how to write clearly, how to give constructive feedback, and how to approach collaboration professionally and thoughtfully.

I would also like to thank Professor Avi Feller for serving on my qualifying exam committee and for being an informal mentor in many ways. He provided thoughtful feedback and suggestions on both research and career decisions, including internships and long-term career choices. His Stat 260 course also greatly deepened my understanding of modern causal inference.

I am deeply grateful to Professor Gokul Upadhyayula for being an outstanding supervisor and collaborator. He gave me the opportunity to work with the Advanced Bioimaging Center of UC Berkeley, one of the world's leading fluorescence microscopy labs. Working with unique biological imaging data and learning from exceptional scientists, including Xiongtao Ruan

and Sayan Seal, was an eye-opening experience. The Biohub conferences consistently exposed me to cutting-edge medical research and broadened my perspective beyond statistics. During the joint project with him, I learned how scientists think, build, and collaborate across disciplines.

I would also like to thank the Yu Group for creating such a diverse, collaborative, and intellectually vibrant environment under Professor Yu's leadership. I am especially grateful to Ana Kenney, who was one of my major mentors during my first year and provided tremendous support and guidance. We worked together on collaborative research with both MD Anderson Cancer Center and the Advanced Bioimaging Center. From her, I learned how to conduct interdisciplinary research and collaborate meaningfully with domain scientists. She has been a wonderful mentor, collaborator, and friend. I would also like to thank Tiffany Tang for her guidance and support during my early years in the group; Aliyah Hsu for being both a wonderful collaborator and friend, and for leading the CD-T project that first introduced me to large language models; Juraj Bodik for collaborating with me on the Crossworld project and exposing me to a more theoretical perspective on causality; and Tao Shi for his support during my internships and career decisions. I also want to thank all the other members of the group whom I had the privilege to learn from and interact with, including Jakob Heiss, Jingfeng Wu, Chengzhong Ye, Austin Zane, Abhinnet Agarwal, Omer Ronen, Corine Elliot, James Duncan, Zeyu Yun, Alex Zhao and Anqi Wang. Learning from this group has been one of the most important parts of my PhD education.

I am also grateful to the causal inference community at Berkeley, especially the Casual Causal Group (Berkeley Causal Lab). I would like to thank Professor Peng Ding and Professor Erin Hartman for their thoughtful feedback on my research. I especially thank Melody Huang for being an exceptional collaborator and mentor on our sensitivity analysis project. She was always supportive and brought many insightful ideas into our collaboration. I also thank Dan Soriano for introducing me to the Berkeley Statistics community and causal inference community, and for his mentorship over the years.

I would like to thank the Department of Statistics at Berkeley for fostering such a diverse and intellectually rich environment. I am especially grateful to my office mates and extended office family: Sizhu, Karissa, Seunghoon, Arisa, Addison, and Lei. We shared happiness, stress, uncertainty, and growth together. They made my life in Berkeley far more colorful and memorable.

Finally, I would like to thank my parents and my family for their unconditional love and support. It is only because of them that I have been brave enough to continue pushing beyond my own boundaries. They have always been my strongest source of comfort, security, and support. I want to thank Yidi Wu, who has been another anchor of my life. Thanks to Chenxin Qiu, who shared so many memories with me and supported me through all the emotions and challenges of building a life in a foreign country. And thanks to all the friends I met through badminton, ballet, and life beyond academia. It is impossible to list every name here, but their importance to me is immeasurable. Loving and being loved has been one of the deepest motivations that carried me through this journey and through life itself.

Contents

Chapter 1

Model Screening and Uncertainty Quantification for Causal Inference

1.1 Introduction

As causal machine learning develops rapidly, a growing number of estimation strategies have been proposed for causal estimands such as the average treatment effect (ATE) and conditional average treatment effect (CATE). Consequently, trustworthy model screening and uncertainty quantification have become increasingly important for ensuring the safety and reliability of causal conclusions, especially in high-stakes domains such as healthcare, public policy and economics. Unlike traditional supervised machine learning problems where outcomes are observed directly and out-of-sample predictive performance provides a natural evaluation criterion, causal estimands are intrinsically unobserved. This makes reliable model assessment more challenging since standard prediction-based validation procedures are not readily applicable.

Existing approaches to causal model screening and uncertainty quantification typically rely on surrogate metrics, nuisance function estimation, or additional structural assumptions. However, these approaches come with their respective limitations. For instance, many asymptotic results rely on consistency or convergence-rate assumptions for nuisance estimators such as propensity scores or outcome regressions, and these assumptions are difficult to verify in practice. On the other hand, assumption-free approaches based on conformal prediction often produce intervals that are too conservative to be practically useful in causal applications.

In this paper, we propose a new framework, PCS-CATE, for treatment effect estimation and uncertainty quantification (UQ) based on the Predictability-Computability-Stability (PCS) framework for veridical data science of Yu and Kumbier [99], Yu and Barter [98], and Rewolinski and Yu [71]. The PCS framework recognizes that data-driven conclusions are the results of a multi-step process of modeling considerations and human judgment calls. It provides a principled way to combine multiple metrics and model screening criteria to

achieve stable and robust predictions. We seek to carry over the PCS framework and the veridical data science principles to the causal inference setting by developing a stable and reliable procedure for causal model screening and uncertainty quantification. This chapter builds on both StaDISC (Dwivedi et al. [24]) and PCS-UQ (Agarwal et al. [3]), which is an extension of Chapter 13 of Yu and Barter [98].

Our contributions are several-fold. First, we extend the PCS-UQ [3] from supervised learning to CATE estimation (PCS-CATE). We replace the standard predictability check used in supervised learning, which is unavailable without the counterfactual ground truth in causal inference, with an aggregation of multiple evaluation metrics, and replace the original calibration step with a subgroup calibration procedure tailored to treatment effect estimation. Second, we re-introduce Cal and WCal as oracle-free subgroup calibration metrics, which compare predicted CATEs against within-bin difference-in-means (or Hájek-IPW) contrasts and provide a finite-sample stand-in for predictive accuracy. Cal was first introduced by Dwivedi et al. [24] as a reality check step; we adopt it as a formal screening metric, and also propose a weighted version WCal for efficiency and stability.

Third, for uncertainty quantification, we propagate data and algorithmic uncertainty via block-bootstrap of the selected estimators and rescale the pooled out-of-bag (OOB) quantile band by a multiplier $\hat{\gamma}$ chosen to satisfy a weighted subgroup-coverage criterion. This is an interval-level calibration step that, to our knowledge, has not previously been used for CATE. Fourth, for working with observational data, we introduce an additional propensity-model screening step combining balance diagnostics, expected calibration error, and a proper score, and propagate propensity-model uncertainty through bootstrap. Finally, through extensive experiments on synthetic and semi-synthetic datasets, we demonstrate that the proposed PCS procedure delivers more stable model screening and near-nominal coverage with modest width inflation in comparison to existing single-metric and per-method baselines. All in all, we provide practical implementation guidelines for both randomized controlled trials and observational studies.

Related work This paper builds on three related strands of literature: model screening and evaluation in causal inference, propensity-score model diagnostics, and the Predictability–Computability–Stability (PCS) framework which includes uncertainty quantification.

Künzel et al. [45] introduce a general framework for estimating CATE with metalearners and baselearners, and propose X-learner which can exploit structural properties of the CATE function and is provably efficient. Nie and Wager [63] propose minimizing the R-loss which depends on residualized outcome and residualized treatment, and prove that the resulting estimator achieves quasi-oracle rates despite estimating nuisance functions. Schuler et al. [85] conduct a horserace of several model screening approaches for treatment effect estimation and find that the R-loss validation metric is the most consistent in selecting a high-performing model, adding support to the claims of Nie and Wager [63] and highlighting the usefulness of the metric in model screening settings where treatment effect model may be fitted by any algorithm. However, two recent empirical studies map the landscape and come to a similar

conclusion that no single screening criterion uniformly outperforms across all settings. Curth and Van Der Schaar [20] argue against any global winner in model screening criteria and characterize the success and failure modes of screening criteria as functions of the data and screening strategies. Mahajan et al. [58] benchmark 34 surrogate metrics against 415 estimators on 78 datasets and find that simple residualized scores perform competitively on average but that no single metric uniformly dominates. We contribute to this literature by developing a stable framework for model screening in treatment effect estimation.

Another related line of work studies calibration for treatment effect estimation. [24] propose StaDISC, a PCS-inspired framework for identifying stable and interpretable subgroups with heterogeneous treatment effects, by evaluating whether predicted subgroup effects are empirically calibrated and stable across perturbations. [50] show that ML CATE estimators can exhibit substantial bias and propose a model-agnostic calibration procedure that aligns predicted treatment effect with subgroup difference-in-means estimates. Our work contributes to this literature by introducing subgroup calibration metrics.

For observational studies, Rubin [83] highlights how propensity scores can be used to help design observational studies and argues that propensity score models should be evaluated based on their ability to balance covariate distributions. Models with large residual imbalance should be rejected. Austin [6] discuss the methods for evaluating whether a propensity score model is correctly specified and propose the use of sampling distribution of standardized difference to assess post-adjustment covariate balance. Our approach extends this by jointly screening propensity-score models using weighted standardized mean difference and several other relevant metrics.

Finally, our work builds on the Predictability–Computability–Stability (PCS) framework for veridical data science first introduced by Yu and Kumbier [99] and further developed in book form by Yu and Barter [98]. Rather than relying solely on a fitted model or asymptotic approximation, PCS focuses on reality checks in model screening and stability across perturbations of the data-analysis pipeline. Regarding uncertainty quantification, PCS perturbation intervals are formalized for regression problems in Yu and Barter [98] and extended to multi-class classification by Agarwal et al. [3]. Our work brings the PCS framework to causal treatment-effect estimation, and replaces the standard prediction-based checks with criteria tailored to treatment effect estimation in both randomized and observational settings.

Outline The remainder of the paper is organized as follows. In Section 1.2, we set up the notations, discuss the steps of our model screening procedure, and elucidate the validation check metrics and details on estimating the confidence interval. Next, in Section 1.3, we extend the framework to observational studies by incorporating an additional propensity score model screening stage and discuss the associated diagnostics for evaluating the propensity score model. Then, we present the complete step-by-step implementation pipeline for the observational setting and highlight the key differences relative to the RCT procedure. To demonstrate the efficacy of our method, in Section 1.4, we evaluate the proposed framework through extensive empirical experiments and present results. Finally, we discuss the

strengths and limitations of the framework, along with directions for future research.

1.2 PCS-CATE Model Screening and UQ in RCT

We first introduce the setup and notation. We consider a randomly sampled dataset $\mathcal{D} = \{(X_i, Z_i, Y_i)\}_{i=1}^n$, where X_i is the covariate vector, $Z_i \in \{0, 1\}$ is the binary treatment indicator (randomly assigned with known probability e in RCTs), and Y_i is the observed outcome. Let $\alpha \in (0, 1)$ denote the desired significance level for uncertainty quantification. Given S candidate CATE algorithms, we write $\hat{\tau}_k(x; \mathcal{D}')$ for the estimated CATE from the k-th estimator fitted on a dataset $\mathcal{D}'$ (e.g., a bootstrap replicate of $\mathcal{D}$). Given M evaluation metrics, we write $L_m(\hat{\tau}_k; \mathcal{D}'')$ for the value of the m-th metric for the k-th estimator on a dataset $\mathcal{D}''$.

The procedure exposes three user-specified hyperparameters that together determine its behaviour: (i) the candidate library $\{\hat{\tau}_k\}_{k=1}^S$ of CATE estimators screened in Step 1, which controls the inductive-bias diversity available to the screening and ensembling stages; (ii) the metrics for screening and retention threshold K in the top-K rank-aggregation step (Section 1.2.1), which trades off ensemble stability against the risk of admitting poorly-fitting methods; and (iii) the number of subgroups L used in the $\hat{\gamma}$-calibration step (Section 1.2.2), which controls the granularity of the calibration check.

We first outline the high-level steps of the procedure; details of the evaluation metrics and the calibration factor are deferred to subsequent sections.

1. **Model Screening ("P" in PCS).**

   a) *Data splitting.* Randomly split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a validation set $\mathcal{D}_{\text{val}}$, block-stratified on Z_i (and on any other blocking variables specified by the RCT design) to keep treatment and control balanced across the split.[1]

   b) *Model fitting.* Train all S candidate CATE estimators on $\mathcal{D}_{\text{train}}$ and evaluate each one on $\mathcal{D}_{\text{val}}$ using all M evaluation metrics (see Section 1.2.1 for details; we use multiple screening metrics rather than relying on a single one).

   c) *Validation check and screening.* For each of the M metrics, rank the S estimators from 1 (best) to S (worst). Aggregate the M ranks for each estimator (e.g., by averaging) to obtain a final rank, and retain the top-K estimators, where K is chosen by the practitioner. This step follows the PCS rank-aggregation procedure, which has been shown to increase both stability and accuracy in real-world applications [94].

2. **Bootstrap and Calibration.**

   a) *Block bootstrap.* Block-bootstrap the full dataset B times, with treated and control units resampled separately. For each replicate $b = 1, \ldots, B$, refit all K re-

[1] Cross-validation may be used when the sample size is small.

tained estimators on $\mathcal{D}^{(b)}$ and record their CATE estimates on the out-of-bag (OOB) indices $\mathcal{I}^{(b)} = \{i : i \notin \mathcal{D}^{(b)}\}$.

b) *OOB aggregation.* For each individual i, pool the OOB estimates across the K retained methods and B replicates,

$$\mathcal{P}_i = \big\{\hat{\tau}_k^{(b)}(X_i) : i \in \mathcal{I}^{(b)},\ k = 1, \ldots, K,\ b = 1, \ldots, B\big\},$$

and define the OOB median and empirical quantiles

$$\hat{q}_{1/2}(X_i) = \text{median}(\mathcal{P}_i), \quad \hat{q}_{\alpha/2}(X_i) = Q_{\alpha/2}(\mathcal{P}_i), \quad \hat{q}_{1-\alpha/2}(X_i) = Q_{1-\alpha/2}(\mathcal{P}_i).$$

c) *Subgroup calibration.* Select $\hat{\gamma} \geq 1$ as described in Section 1.2.2.

This step extends the bootstrap-and-calibration step of PCS-UQ [3] to the causal setting. The extension is non-trivial: the CATE is unobserved, so in place of a held-out target we use the subgroup difference-in-means (DiM) as a noisy but unbiased estimator of the subgroup ATE. As a consequence, we cannot deliver the finite-sample coverage guarantees that conformal prediction or PCS-UQ enjoy. Instead, the calibration step serves as a necessity check, triggering interval inflation whenever the OOB intervals are demonstrably miscalibrated against the subgroup contrasts (see Section 1.2.2 for the formal argument).

3. **Point and Interval Estimation.**

a) *Point estimate.* Refit each of the K retained estimators once on the full dataset $\mathcal{D}$. The point estimate at x is the mean of their predictions,

$$\hat{\tau}(x) = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_k(x;\, \mathcal{D}).$$

b) *Calibrated interval.* For each individual i, the calibrated interval is constructed by inflating the OOB quantile band around the median by $\hat{\gamma}$:

$$I_i = \Big[\hat{q}_{1/2}(X_i) - \hat{\gamma}\big(\hat{q}_{1/2}(X_i) - \hat{q}_{\alpha/2}(X_i)\big),\ \ \hat{q}_{1/2}(X_i) + \hat{\gamma}\big(\hat{q}_{1-\alpha/2}(X_i) - \hat{q}_{1/2}(X_i)\big)\Big].$$

1.2.1 Validation Check Metrics

We use four metrics to rank candidate estimators on the validation set $\mathcal{D}_{\text{val}}$: two variants of a subgroup calibration score (Cal, WCal) and two variants of the R-risk (Rrisk, Rrisk_raw). All four return non-negative scores, with smaller values indicating better fit. For each metric we rank the candidate estimators from 1 (best) to S (worst); the four ranks are then averaged to obtain a single screening score, and the $K = 3$ estimators with the smallest average rank are retained. This rank-aggregation step follows the PCS rank-aggregation procedure of

Tang et al. [94], which has been shown to improve both stability and accuracy over single-metric selection in real-world applications. Of the four metrics, WCal and Rrisk_raw are introduced in this paper; Cal follows Dwivedi et al. [24] but is repurposed here as a formal screening criterion (in the original work it served only as a reality check), and Rrisk is taken directly from Nie and Wager [63], where it is shown to be quasi-oracle efficient under standard regularity conditions.

Calibration score (Cal). For each candidate estimator $\hat{\tau}$, form $Q = 5$ quantile bins $\{G_j\}_{j=1}^{Q}$ from the training-set predictions $\hat{\tau}(X_i)$, $i \in \mathcal{D}_{\text{train}}$, and apply the cutpoints to $\mathcal{D}_{\text{val}}$. Within bin j, let $\bar{M}_j = |G_j|^{-1} \sum_{i \in G_j} \hat{\tau}(X_i)$ denote the mean predicted CATE and $\hat{\tau}_j^{\text{DiM}} = |G_j^1|^{-1} \sum_{i \in G_j, Z_i = 1} Y_i - |G_j^0|^{-1} \sum_{i \in G_j, Z_i = 0} Y_i$ the within-bin difference-in-means estimate of the subgroup ATE. The calibration score follows Dwivedi et al. [24],

$$\text{Cal}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \sum_{j=1}^{Q} \frac{|G_j|}{|\mathcal{D}_{\text{val}}|} \left| \bar{M}_j - \hat{\tau}_j^{\text{DiM}} \right|,$$

and rewards estimators whose predicted heterogeneity aligns with observed subgroup contrasts. In its original form, Cal (and the normalised R_C^2 variant of Dwivedi et al. [24]) was used as a reality check rather than a screening criterion: on the VIGOR data they study, none of the 18 candidate estimators passed the check, and estimator screening was carried out separately through a t-statistic-based procedure on the most extreme quantile bin. We instead repurpose Cal as a formal screening metric, treating a low value as direct evidence of good subgroup-level fit.

Variance-weighted calibration score (WCal). Cal weights all bins by size, but the precision of $\hat{\tau}_j^{\text{DiM}}$ varies considerably across bins. To stabilise the score against this source of noise, we introduce a variance-weighted variant that scales each bin's contribution by the inverse Neyman variance of the DiM estimator,

$$\text{WCal}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \sum_{j=1}^{Q} \frac{|G_j|}{|\mathcal{D}_{\text{val}}|} \cdot \frac{\left(\bar{M}_j - \hat{\tau}_j^{\text{DiM}}\right)^2}{\widehat{\text{Var}}(\hat{\tau}_j^{\text{DiM}})},$$

where $\widehat{\text{Var}}(\hat{\tau}_j^{\text{DiM}}) = s_{1,j}^2/|G_j^1| + s_{0,j}^2/|G_j^0|$. WCal penalises miscalibration more heavily in bins where the empirical signal is reliable, and discounts bins where DiM noise would otherwise dominate the unweighted score.

R-risk variants. The R-risk is the empirical plug-in of the R-learner loss [63], which residualises the outcome against an estimated main-effect surface and the treatment against its known propensity:

$$\text{Rrisk}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} \left[(Y_i - \hat{m}(X_i)) - (Z_i - e)\hat{\tau}(X_i) \right]^2,$$

where $\hat{m}(X) \approx \mathbb{E}[Y \mid X]$ is a cross-fitted estimate of the main-effect surface obtained on $\mathcal{D}_{\text{train}}$. Although Rrisk inherits the rate properties of double-residualisation, its empirical reliability as a screening criterion depends on the quality of $\hat{m}$, and recent benchmarking work has flagged this dependence as a practical concern [58, 20]. We therefore also include a nuisance-free reference,

$$\text{Rrisk\_raw}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} \big[Y_i - (Z_i - e)\hat{\tau}(X_i)\big]^2,$$

obtained by setting $\hat{m} \equiv 0$. Rrisk_raw is noisier than the nuisance-adjusted variant when the baseline signal is strong but is immune to the failure modes induced by a poorly estimated $\hat{m}$; including both allows the rank-aggregation step to balance the two failure modes rather than committing to one variant a priori.

1.2.2 Choosing $\hat{\gamma}$

This section explains the method for choosing $\hat{\gamma}$ which determines the width of the confidence interval. Let $\hat{q}_{1/2}(X_i)$ be the OOB median estimates. We first partition the data into L subgroups $\{G_\ell\}_{\ell=1}^{L}$ based on quantiles of $\hat{q}_{1/2}(X_i)$. Within each subgroup, compute the empirical treatment effect

$$\hat{\Delta}_\ell = \frac{1}{n_{1,\ell}} \sum_{i \in G_\ell, Z_i = 1} Y_i - \frac{1}{n_{0,\ell}} \sum_{i \in G_\ell, Z_i = 0} Y_i,$$

with estimated variance

$$\widehat{\text{Var}}(\hat{\Delta}_\ell) = \frac{s_{1,\ell}^2}{n_{1,\ell}} + \frac{s_{0,\ell}^2}{n_{0,\ell}},$$

where $s_{1,\ell}^2$ and $s_{0,\ell}^2$ are the sample variances in the treated and control groups within G_ℓ.

Next, we define the aggregated interval

$$\bar{I}_\ell(\gamma) = \left[\frac{1}{|G_\ell|} \sum_{i \in G_\ell} l_i(\gamma),\ \frac{1}{|G_\ell|} \sum_{i \in G_\ell} u_i(\gamma) \right],$$

and let $w_\ell = \left(\widehat{\text{Var}}(\hat{\Delta}_\ell) + \varepsilon \right)^{-1}$ for a small constant $\varepsilon > 0$. We select $\hat{\gamma}$ as the smallest value such that

$$\frac{\sum_{\ell=1}^{L} w_\ell \, \mathbf{1}\{\hat{\Delta}_\ell \in \bar{I}_\ell(\gamma)\}}{\sum_{\ell=1}^{L} w_\ell} \geq 1 - \eta,$$

for a tolerance level $\eta \in (0, 1)$. In practice we choose L so that each subgroup contains at least $n_{\min}$ treated and $n_{\min}$ control units, with $n_{\min} = 50$ as a default; this caps L at roughly $n_{\text{train}}/(2n_{\min})$. A smaller $n_{\min}$ produces more subgroups and a finer calibration check at

the cost of noisier within-subgroup DiM contrasts. Because the weighted subgroup-coverage criterion is discrete when L is small, we define

$$L_{\text{eff}} = \frac{(\sum_{\ell=1}^{L} w_\ell)^2}{\sum_{\ell=1}^{L} w_\ell^2},$$

and use $\eta = \max\{\alpha, 1/L_{\text{eff}}\}$. Thus, when many effective subgroups are available, the calibration tolerance is aligned with the nominal interval level. When the number of effective subgroups is small, the rule allows approximately one effective subgroup miss rather than requiring all subgroups to be covered.

The subgroup calibration step should be interpreted as calibration for subgroup-average effects rather than as distribution-free individual-level coverage. Under randomization, $\hat{\Delta}_\ell$ is an unbiased estimator of the subgroup-average treatment effect $\Delta_\ell = |G_\ell|^{-1} \sum_{i \in G_\ell} \tau(X_i)$, with variance estimated by the Neyman variance estimator. Hence, if $\hat{\Delta}_\ell$ lies in the aggregated interval $\bar{I}_\ell(\gamma)$ with slack larger than its estimation noise, then Δ_ℓ is covered with high probability. Conversely, persistent subgroup-level bias or under-dispersion in the individual intervals will cause $\hat{\Delta}_\ell$ to fall outside $\bar{I}_\ell(\gamma)$ with high probability as subgroup size grows. Thus, the calibration step provides an empirical finite-sample correction for subgroup-level miscoverage, although it does not imply distribution-free individual-level coverage in the conformal sense.

1.3 Extending PCS-CATE to Observational Studies

When working with observational data, the model screening and uncertainty quantification steps are similar to the CATE under RCT case in section 1.2, but observational studies require an additional nuisance-estimation layer for propensity score. Most CATE estimators for observational data rely on an estimated propensity score or balancing score. Since this nuisance component directly affects both CATE model screening and final estimation, we screen propensity score models before fitting the final CATE estimators and propagate the resulting propensity-model uncertainty in the final interval construction.

1.3.1 Propensity score model screening

Existing propensity-score diagnostics emphasize either covariate balance, corresponding to the design-based view, or nuisance convergence rates, corresponding to the asymptotic semiparametric view. Both are useful but incomplete in finite samples. On one hand, balance diagnostics are often restricted to low-order moments, especially mean balance, and can miss nonlinear or distributional imbalance. On the other hand, convergence-rate conditions are not directly observable in real data. We therefore combine balancing diagnostics with probabilistic diagnostics as finite-sample proxies for propensity-model quality. The goal is not to certify that $\hat{e}(x)$ is the true propensity score, but to select models that jointly provide reasonable balance and probabilistic fit.

Let $\hat{e}_k(X_i)$ denote the out-of-fold propensity score prediction from candidate model k. We truncate the propensity score estimates to lie in $[0.01, 0.99]$ and then compute weighting-based diagnostics. We first define

$$w_{ik} = \frac{Z_i}{\hat{e}_k(X_i)} + \frac{1 - Z_i}{1 - \hat{e}_k(X_i)}.$$

For each candidate model, we compute the following diagnostics.

Mean balance For each covariate X_j, compute the weighted standardized mean difference

$$\mathrm{SMD}_{jk} = \frac{\bar{X}_{1j}^{w,k} - \bar{X}_{0j}^{w,k}}{s_j},$$

where $\bar{X}_{1j}^{w,k}$ and $\bar{X}_{0j}^{w,k}$ are the weighted treated and control means under model k, and s_j is the pooled pre-weighting standard deviation. We summarize mean imbalance by $\max_j |\mathrm{SMD}_{jk}|$.

Distributional balance To check balance beyond means, we compute a weighted Kolmogorov–Smirnov discrepancy for each covariate:

$$\mathrm{KS}_{jk} = \sup_x \left| \hat{F}_{1j}^{w,k}(x) - \hat{F}_{0j}^{w,k}(x) \right|,$$

where

$$\hat{F}_{1j}^{w,k}(x) = \frac{\sum_{i:Z_i=1} w_{ik} \mathbf{1}\{X_{ij} \le x\}}{\sum_{i:Z_i=1} w_{ik}}, \qquad \hat{F}_{0j}^{w,k}(x) = \frac{\sum_{i:Z_i=0} w_{ik} \mathbf{1}\{X_{ij} \le x\}}{\sum_{i:Z_i=0} w_{ik}}.$$

We summarize distributional imbalance by the average KS statistic.

Calibration We partition observations into bins $\{B_b\}_{b=1}^{B_e}$ according to quantiles of $\hat{e}_k(X_i)$. Let

$$\bar{e}_{bk} = |B_b|^{-1} \sum_{i \in B_b} \hat{e}_k(X_i), \qquad \bar{Z}_b = |B_b|^{-1} \sum_{i \in B_b} Z_i.$$

The binning-based expected calibration error (ECE) is

$$\mathrm{ECE}_k = \sum_{b=1}^{B_e} \frac{|B_b|}{n} |\bar{Z}_b - \bar{e}_{bk}|.$$

Proper scoring rule We evaluate the out-of-fold log loss

$$\text{LogLoss}_k = -\frac{1}{n}\sum_{i=1}^{n}\left\{Z_i \log \hat{e}_k(X_i) + (1 - Z_i)\log[1 - \hat{e}_k(X_i)]\right\}.$$

We aggregate these diagnostics using constrained average ranking. Since the propensity score plays a crucial role, we suggest an initial screening step based on visual inspection of balance and calibration diagnostics before applying the aggregated ranking. Among the remaining candidates, we rank models separately by each metric, with rank 1 corresponding to the best value, and use the average rank as the final screening score. We retain the top K_e propensity score models and aggregate the retained models into a single propensity score estimate by taking the mean. To reflect propensity-model uncertainty, downstream CATE estimation is repeated across the retained propensity models.

1.3.2 Step-by-step implementation

We use the same notation as in Section 1.2.1, with the following additions. We let $\hat{e}_k(X_i)$ denote the out-of-fold propensity score from candidate model k. K_e denotes the number of retained propensity models. $\hat{e}(X_i) = K_e^{-1}\sum_{k=1}^{K_e}\hat{e}_k(X_i)$ is the propensity ensemble used in downstream estimation.

1. **Propensity score model screening**

   a) Cross-fitting: For each candidate propensity model k, obtain out-of-fold predictions $\hat{e}_k(X_i)$ via V-fold cross-fitting on the full data. Truncate all estimates to $[0.01, 0.99]$ before computing any weighting-based diagnostic.

   b) Diagnostics: Compute the four diagnostics defined above—$\max_j |\text{SMD}_{jk}|$, mean KS_k, ECE_k, LogLoss_k—for each candidate model k.

   c) Visual screening and ranking: Inspect covariate-balance and calibration plots; eliminate models with extreme imbalance (e.g., $\max_j |\text{SMD}_{jk}| > 0.1$) before ranking. Among the remaining candidates, rank each model by each diagnostic (rank 1 = best) and average the ranks. Retain the top K_e propensity models.

   d) Propensity ensemble: Aggregate into a single score:

$$\hat{e}(X_i) = \frac{1}{K_e}\sum_{k=1}^{K_e}\hat{e}_k(X_i), \qquad \text{truncated to } [0.01, 0.99].$$

2. **CATE model screening**

   a) Data splitting: Randomly split the data into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$, block-stratified on Z_i.

b) Nuisance pre-computation: On $\mathcal{D}_{\text{train}}$: (i) cross-fit the outcome nuisance $\hat{m}(X)$ for use in Rrisk; (ii) re-fit the propensity ensemble $\hat{e}(\cdot)$ and apply it to $\mathcal{D}_{\text{val}}$.

c) CATE fitting: Train all candidate CATE estimators on $\mathcal{D}_{\text{train}}$ using $\hat{e}(\cdot)$.

d) Metric computation: For each estimator $\hat{\tau}_k$, compute the following oracle-free metrics on $\mathcal{D}_{\text{val}}$. First, we form bins $\{G_j\}_{j=1}^K$ from quantiles of the training-set predictions $\hat{\tau}_k(X_i)$, $i \in \mathcal{D}_{\text{train}}$, applied to $\mathcal{D}_{\text{val}}$. Let $\bar{M}_j = |G_j|^{-1} \sum_{i \in G_j} \hat{\tau}(X_i)$ be the mean predicted CATE in bin j, and let

$$\hat{\Delta}_j^{\text{IPW}} = \frac{\sum_{i \in G_j} Z_i Y_i / \hat{e}(X_i)}{\sum_{i \in G_j} Z_i / \hat{e}(X_i)} - \frac{\sum_{i \in G_j} (1 - Z_i) Y_i / [1 - \hat{e}(X_i)]}{\sum_{i \in G_j} (1 - Z_i) / [1 - \hat{e}(X_i)]}$$

be the Hájek IPW contrast in bin j.

- **Cal**$^{\text{IPW}}$: replace the DiM contrast in Cal by $\hat{\Delta}_j^{\text{IPW}}$:

$$\text{Cal}^{\text{IPW}}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \sum_{j=1}^{K} \frac{|G_j|}{|\mathcal{D}_{\text{val}}|} \big| \bar{M}_j - \hat{\Delta}_j^{\text{IPW}} \big|.$$

- **WCal**$^{\text{IPW}}$: weight each bin by the precision of $\hat{\Delta}_j^{\text{IPW}}$:

$$\text{WCal}^{\text{IPW}}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \sum_{j=1}^{K} \frac{|G_j|}{|\mathcal{D}_{\text{val}}|} \cdot \frac{\big(\bar{M}_j - \hat{\Delta}_j^{\text{IPW}}\big)^2}{\widehat{\text{Var}}(\hat{\Delta}_j^{\text{IPW}})},$$

where $\widehat{\text{Var}}(\hat{\Delta}_j^{\text{IPW}})$ is the estimated variance of the Hájek IPW contrast in bin j.

- **R-risk**$^{\text{obs}}$: replace the known propensity e by the individual estimate $\hat{e}(X_i)$:

$$\text{Rrisk}^{\text{obs}}(\hat{\tau};\, \mathcal{D}_{\text{val}}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} \Big[\big(Y_i - \hat{m}(X_i)\big) - \big(Z_i - \hat{e}(X_i)\big) \hat{\tau}(X_i) \Big]^2.$$

- **R-risk**$^{\text{obs}}_{\text{raw}}$: same without outcome nuisance ($\hat{m}(X) \equiv 0$).

e) Aggregated screening: Rank estimators by each metric and average the ranks. Retain the top-K CATE estimators.

3. **Bootstrap and calibration**

a) Block-bootstrap with propensity propagation: Block-bootstrap the full data B times. In each replicate b:

(i) Re-fit the K_e retained propensity models on $\mathcal{D}^{(b)}$; form the OOB propensity ensemble $\hat{e}^b(X_i)$ for the OOB indices $\mathcal{I}^{(b)}$.

(ii) Fit the K retained CATE estimators on $\mathcal{D}^{(b)}$ using $\hat{e}^b(\cdot)$; record OOB predictions $\hat{\tau}_k^b(X_i)$ for $i \in \mathcal{I}^{(b)}$.

b) OOB aggregation: For each unit i, collect

$$\mathcal{P}_i = \left\{\hat{\tau}_k^b(X_i) : i \in \mathcal{I}^{(b)},\ k = 1, \ldots, K,\ b = 1, \ldots, B\right\},$$

and form the OOB median and empirical quantiles $\hat{q}_{1/2}(X_i)$, $\hat{q}_{\alpha/2}(X_i)$, $\hat{q}_{1-\alpha/2}(X_i)$ as in the RCT procedure.

c) Subgroup calibration: Partition units into L subgroups by quantiles of $\hat{q}_{1/2}(X_i)$, using a larger minimum subgroup size than in the RCT case to keep IPW variance under control. For each subgroup G_ℓ, compute $\hat{\Delta}_\ell^{\mathrm{IPW}}$ using the full-data propensity ensemble $\hat{e}(\cdot)$. Select $\hat{\gamma}$ via the weighted subgroup-coverage criterion of Section 1.2.2, with $\hat{\Delta}_\ell$ replaced by $\hat{\Delta}_\ell^{\mathrm{IPW}}$ and weights $w_\ell = (\widehat{\mathrm{Var}}(\hat{\Delta}_\ell^{\mathrm{IPW}}) + \varepsilon)^{-1}$.

4. **Point and interval estimation**

a) Point estimate: Re-fit the K_e propensity models and the K CATE estimators on the full data $\mathcal{D}$. The point estimate at x is:

$$\hat{\tau}(x) = \frac{1}{K}\sum_{k=1}^{K} \hat{\tau}_k(x;\, \mathcal{D}).$$

b) Calibrated interval: For each observed unit i, construct the interval from the OOB quantiles exactly as in the RCT procedure:

$$I_i = \Big[\hat{q}_{1/2}(X_i) - \hat{\gamma}\left(\hat{q}_{1/2}(X_i) - \hat{q}_{\alpha/2}(X_i)\right),\ \ \hat{q}_{1/2}(X_i) + \hat{\gamma}\left(\hat{q}_{1-\alpha/2}(X_i) - \hat{q}_{1/2}(X_i)\right)\Big].$$

Key differences from the RCT procedure. The observational procedure differs from its RCT counterpart in three ways, each motivated by the introduction of confounding and the resulting need to estimate nuisance components from the data.

First, an explicit propensity-screening step (Step 1) precedes CATE screening, producing a propensity ensemble $\hat{e}(\cdot)$ that replaces the known e throughout. The propensity score is the foundation on which most modern CATE estimators rest, as DR-learners, R-learners, and IPW-based estimators all consume it as a nuisance input, so a misspecified propensity contaminates everything downstream. Rather than committing to a single specification, we apply the same PCS logic used in the CATE step: rank candidate propensity models by complementary diagnostics (balance diagnostics for the design-based view, calibration and log-loss for the distributional view), retain the top K_e by average rank, and aggregate them into an ensemble. The result is a nuisance estimator that is itself robust to model choice.

Second, all subgroup calibration targets use the Hájek IPW contrast $\hat{\Delta}_\ell^{\mathrm{IPW}}$ in place of the unconfounded DiM. Under randomization the within-subgroup DiM is unbiased for the subgroup ATE; under confounding it is not, and calibrating to a biased target would simply lock in the confounding. The Hájek IPW contrast restores consistency for Δ_ℓ under strong overlap and consistent propensity estimation; we use the Hájek (self-normalized) form rather than the

augmented IPW alternative to avoid introducing an additional outcome-regression nuisance into the calibration target, and prefer it over Horvitz-Thompson because self-normalization stabilizes the contrast in finite samples where extreme weights are common.

Third, propensity-model uncertainty is explicitly propagated through the bootstrap by re-fitting the propensity ensemble in each replicate (Step 3.a(i)), so that the final OOB intervals reflect both data and propensity-model variation. Treating $\hat{e}(\cdot)$ as fixed after the initial fit would understate uncertainty, because the propensity ensemble is itself a function of the data and its error feeds into every downstream CATE estimate. Re-fitting in each replicate captures this error in the natural place, alongside the data and CATE-method variability already being propagated, consistent with the PCS principle that all consequential modelling choices should be perturbed when assessing stability, and yields intervals that honestly reflect all three sources of uncertainty.

1.4 Empirical results and discussion

We evaluate the PCS procedure in three experimental settings. The first uses the four synthetic setups (A–D) from the R-learner paper of Nie and Wager [63], modified to use balanced treatment assignment $e \equiv 0.5$; these vary the complexity of the CATE function and the baseline outcome surface across $n_{\text{train}} \in \{500, 1{,}000, 2{,}000, 5{,}000\}$. The second uses three semi-synthetic data-generating processes from the RealCause benchmark of Neal, Huang, and Raghupathi [60] (LaLonde/CPS, LaLonde/PSID, Twins), in which generative models fitted to real datasets supply both potential outcomes so that the true CATE is known; we re-randomise treatment assignment in each replicate to obtain a balanced RCT. The third reuses the four R-learner setups but retains their original observational propensity scores, evaluating the procedure under confounding. The candidate library is fixed across all settings and consists of 14 estimators: 12 meta-learner variants (S/T/X/R-learner $\times$ lasso/ridge/forest) together with `causal_forest` and `bart`.

Comparisons are made on three axes. For *model screening*, we report the Spearman rank correlation between each screening criterion and the oracle MSE ranking, comparing the average-rank aggregator against each of the four metrics used in isolation. For *point estimation*, we compare the refit ensemble (mean of the $K = 3$ retained estimators refitted on the full training data) against the bootstrap ensemble (median of all $B \times K$ OOB predictions) and against the individual rank-1, rank-2, and rank-3 refits, evaluated by test-set MSE against the true $\tau(X)$. For *uncertainty quantification*, we compare the calibrated PCS pooled interval against the per-rank bootstrap intervals at nominal level $\alpha = 0.05$, reporting test-set coverage and mean width. Simulation details are provided in the appendix.

Simulation pipeline. Each replication proceeds as follows: (i) draw fresh training and test samples from the DGP (re-randomising treatment for the semi-synthetic case); (ii) split the training data 75/25, block-stratified on Z, into screening-train and validation sets; (iii) fit all 14 candidate CATE estimators on the screening-train and rank them on the validation

set using the four oracle-free metrics, retaining the top $K = 3$ by average rank; (iv) draw $B = 500$ block-bootstrap replicates of the full training set, refit the K retained methods on each replicate, and record OOB predictions; (v) calibrate $\hat{\gamma}$ via the weighted subgroup-coverage criterion on the bootstrap medians; (vi) refit the K retained methods on the full training set to form the refit-ensemble point estimate, and form the calibrated PCS pooled interval from the OOB quantile distribution. We report (a) screening quality via Spearman rank correlation against the oracle MSE ranking; (b) test-set MSE of point estimators against the true $\tau(X_i)$; (c) test-set coverage and width of the calibrated PCS pooled interval and the per-rank bootstrap intervals at nominal level $\alpha = 0.05$. The synthetic setups vary the training sample size $n_{\text{train}} \in \{500, 1000, 2000, 5000\}$ with 100 replications per configuration; the semi-synthetic setting uses $n_{\text{train}} = 1000$ with 100 replications.

For observational data, we re-use the four DGP setups (A–D) of Nie and Wager [63] but, unlike the RCT case, retain the original observational propensity scores so that treatment assignment is confounded by X. Each replication follows the same pipeline as the RCT simulation, with two additions: before CATE screening, six candidate propensity models are cross-fitted, ranked by the four diagnostics in Section 1.2.1, and the top $K_e = 2$ are aggregated into the propensity ensemble $\hat{e}(\cdot)$; and every subgroup contrast used for screening and calibration is replaced by its Hájek IPW counterpart. We use $n_{\text{train}} = 2000$ with 100 replications, and test-set MSE is computed against the true $\tau(X_i)$.

Synthetic data (RCT). Figure 1.1 reports the Spearman rank correlation between each screening metric and the oracle MSE ranking. Rrisk and Rrisk_raw track the oracle ranking well in Setups A and D, but their performance degrades as sample size increases in Setup B (nonlinear CATE) and Setup C (constant CATE), pointing to sensitivity to the quality of nuisance estimation. WCal slightly outperforms Cal in all setups except Setup C; both metrics use a fixed five-subgroup partition, and a more adaptive choice would likely improve them further. The average-rank aggregator is rarely the single best metric on any given configuration, but it is consistently among the strongest, delivering stable performance across both sample sizes and setups.

Test-set MSE for the point estimators is shown in Figure 1.2. The refit ensemble (mean of the K retained methods refitted on the full training data) outperforms both the bootstrap ensemble (median of all $B \times K$ OOB predictions) and the individual rank-1/2/3 refits in most configurations. Pooling across the four setups and sample sizes (1,600 replications), averaging the three refits reduces test-set MSE by a mean of 30.4% relative to the mean MSE of the individual refits. The gain is largest in Setup C (47.4%), where the constant-CATE oracle is recoverable by averaging, and in Setup B (29.7%); it remains substantial in Setup A (23.9%) and Setup D (20.8%).

Figures 1.3–1.5 report the test-set coverage and width of the calibrated PCS pooled interval and the per-rank bootstrap intervals, together with the calibration factor $\hat{\gamma}$. Coverage of the pooled interval stays close to the nominal 95% level across sample sizes and setups, with the exception of Setup D at small n_{train}; the per-rank bootstrap intervals are substantially

less reliable. Only 4.5% of replications produce a pooled coverage below 80%, almost all from Setup D (16.5%), with Setups A, B, and C each at or below 1.3%. The per-rank intervals, by contrast, fall below 80% coverage in 24.7%, 33.2%, and 33.5% of replications for ranks 1–3 respectively, with under-coverage particularly severe in Setup D (43.5%, 84.3%, 90.0%). The pooled interval is on average 33.9% wider than the mean per-rank interval (Setup A: 28.2%; B: 29.5%; C: 42.0%; D: 35.9%), a modest premium that buys the substantially improved coverage. The largest width inflation occurs in Setup C, where $\hat{\gamma}$ is itself elevated: although the constant CATE is easy to estimate, the within-bin DiM contrasts are noisy, and the calibration step responds by inflating the interval. Outside Setup C, $\hat{\gamma}$ remains close to one.

Semi-synthetic data (RCT). Screening behaviour on the three RealCause DGPs is shown in Figure 1.6. Rrisk performs particularly well here, likely because the semi-synthetic structure permits accurate nuisance estimation. Cal and WCal again use the default five-subgroup partition and would likely benefit from a finer choice on these larger DGPs. As in the synthetic case, the average-rank aggregator is consistently strong across DGPs without ever being the single best metric on any one of them.

Test-set MSE varies little across estimators, with the refit ensemble performing consistently well (Figure 1.7). Averaging the three refits reduces test-set MSE by a mean of 34.9% relative to the mean of the individual refits, with the largest gain on `lalonde_cps` (43.9%) and the smallest on `twins` (29.2%).

Figures 1.8–1.10 summarise the uncertainty quantification. The pooled interval delivers good coverage without being excessively wide: only 2% of replications produce a pooled coverage below 80% (0% on CPS and Twins, 6% on PSID), compared with 20.7%, 26.0%, and 24.7% of replications for the rank-1, rank-2, and rank-3 bootstrap intervals respectively, with under-coverage concentrated on `lalonde_cps` (37–42%). The pooled interval is on average 33.4% wider than the mean per-rank interval, but the premium varies considerably by DGP: it is largest on `lalonde_cps` (55.2%), where bias from a single method is most pronounced, falls to 36.5% on PSID, and is almost negligible on `twins` (8.6%), where the candidate methods agree closely.

Synthetic data (observational). Point estimation results on the observational simulations are shown in Figure 1.11. Averaging the three refits reduces test-set MSE by a mean of 29.3% relative to the mean MSE of the individual refits (Setup A: 24.3%; B: 31.1%; C: 43.9%; D: 17.8%). The pattern across setups mirrors the RCT case: the constant-CATE Setup C benefits most from averaging, while Setup D, with its sign-changing CATE that shares structure with the baseline, benefits least.

Uncertainty quantification results are shown in Figures 1.12 and 1.13. Coverage of the PCS pooled interval is generally close to nominal (mean $\approx$ 92%) but degrades on Setup D (87.1%), where IPW noise compounds with the signal-separation challenge of the DGP. Across the four setups, 4–20% of replications produce a pooled coverage below 80% (A: 4%; B: 3%; C: 6%; D: 20%), whereas the per-rank intervals fall below 80% much more frequently:

33–50% for rank-1, 33–59% for rank-2, and 34–81% for rank-3, with the worst rates again on Setup D. The pooled interval is on average 41.7% wider than the mean per-rank interval (A: 41.8%; B: 30.5%; C: 67.5%; D: 26.6%). The larger width premium relative to the RCT case reflects the additional propensity-model variability propagated through the bootstrap, together with the wider IPW subgroup contrasts that drive $\hat{\gamma}$ upward.

Result Summary The proposed procedure outperforms its single-metric and single-method baselines on each of the three evaluation axes and across all three experimental settings. For *model screening*, the average-rank aggregator is consistently among the strongest criteria across DGPs and sample sizes, even though no single underlying metric dominates uniformly. For *point estimation*, the refit ensemble of the $K = 3$ retained estimators reduces test-set MSE by roughly 30% relative to the mean of the individual refits (30.4% on the synthetic RCT setups, 34.9% on the RealCause DGPs, and 29.3% on the observational simulations), with the largest gains in configurations where the candidate methods disagree most. For *uncertainty quantification*, the calibrated PCS pooled interval achieves coverage close to the nominal 95% level in nearly all configurations, falling below 80% in only 4.5%, 2%, and roughly 8% of replications across the three settings, while the corresponding per-rank bootstrap intervals fall below 80% in 24–33%, 20–26%, and 33–81% of replications respectively. The pooled interval is on average 33–42% wider than the mean per-rank interval, a modest premium for the substantially improved coverage.

Discussion Estimating heterogeneous treatment effects is a hard problem: counterfactual outcomes are unobserved, so no single oracle-free metric can be trusted for model screening, and with the proliferation of CATE estimators carrying different inductive biases, no single method is uniformly best across DGPs. The PCS procedure addresses both sources of fragility by aggregating across metrics during screening and across methods and bootstrap replicates during estimation and uncertainty quantification. The intuition is simple: *different metrics catch different failure modes, and different estimators perform well in different regions of covariate space and extrapolate differently outside the training support*; pooling them yields a more stable and reliable picture of the CATE than any single choice. Across the synthetic, semi-synthetic, and observational simulations reported above, this aggregation pays off, delivering stable model screening, roughly 30% lower point-estimation error than the average individual refit, and pooled intervals close to nominal 95% coverage at a modest width premium.

We are currently extending the empirical evaluation to a broader range of synthetic and real-world benchmarks, and adapting the PCS philosophy to additional causal estimands.

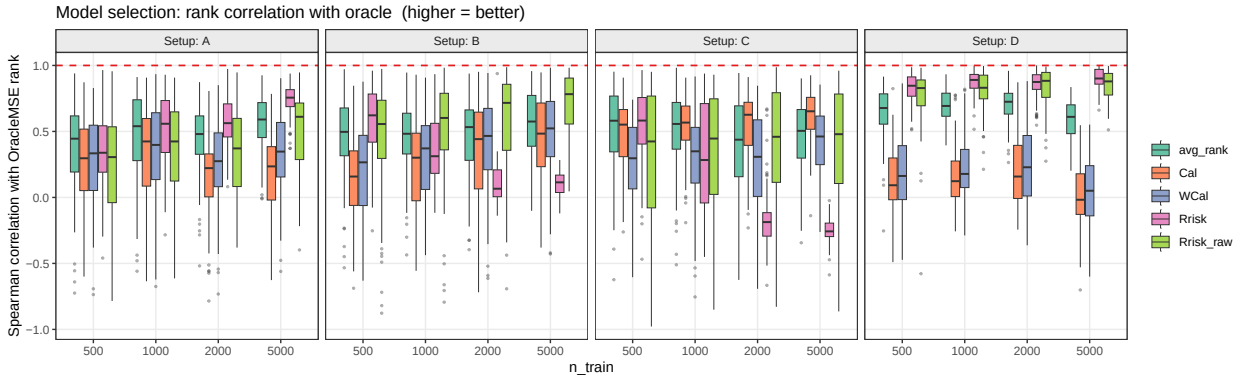

Figure 1.1: Spearman rank correlation between each screening metric (avg_rank, Cal, WCal, Rrisk, Rrisk_raw) and the oracle MSE ranking, across training sample sizes $n_{\text{train}} \in \{500, 1000, 2000, 5000\}$ and DGP setups A–D (100 replications each). Higher is better.

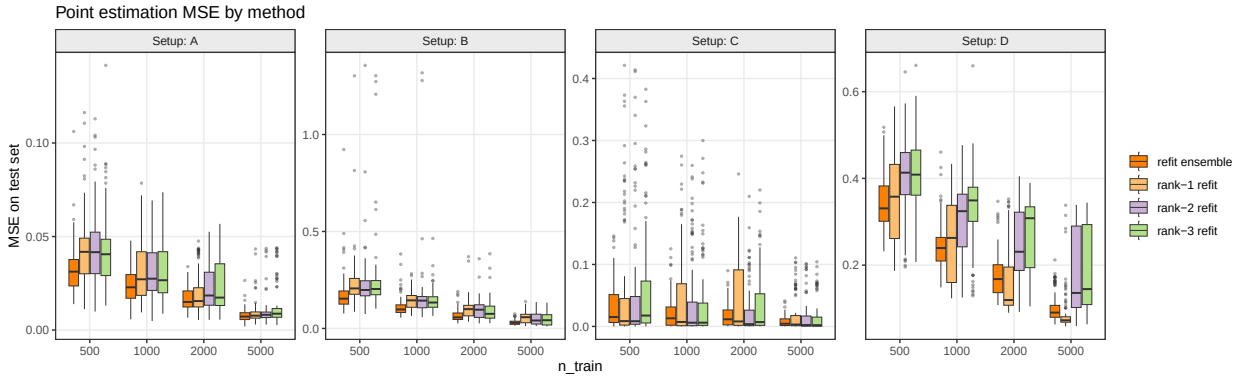

Figure 1.2: Test-set MSE of point estimators across sample sizes and setups A–D: bootstrap ensemble (median of all $B \times K$ OOB predictions), refit ensemble (mean of K methods refitted on full training data), and individual rank-1/2/3 refits. Lower is better.

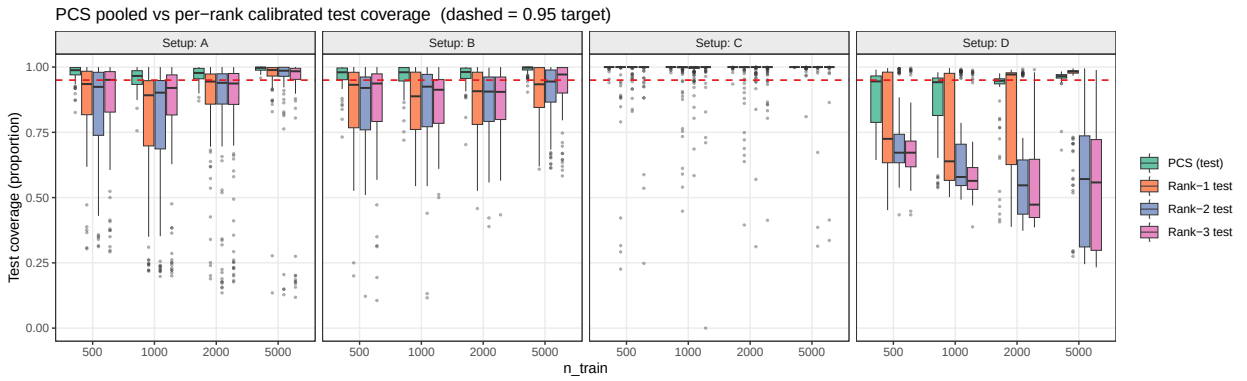

Figure 1.3: Test-set coverage of the PCS pooled interval and per-rank calibrated intervals (rank-1, rank-2, rank-3) across sample sizes and setups A–D. Dashed line at the nominal 95% level.

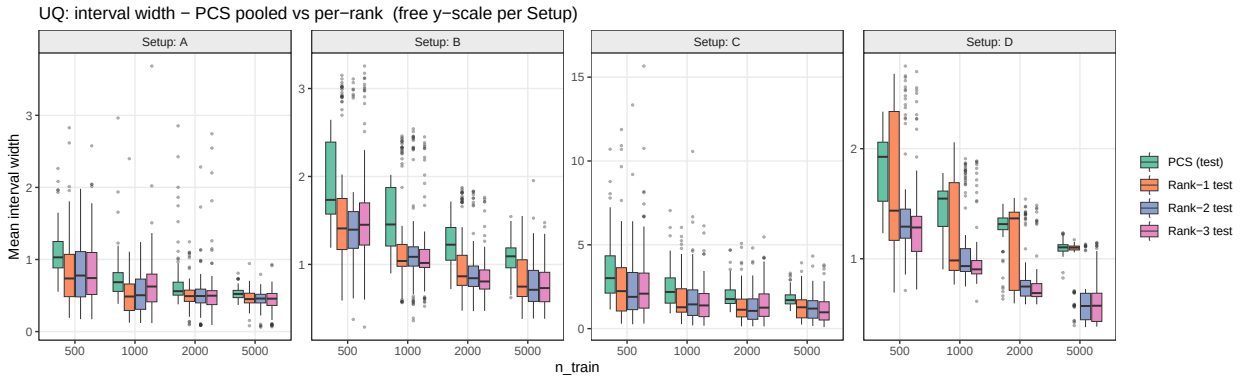

Figure 1.4: Mean interval width of the PCS pooled and per-rank calibrated intervals across sample sizes and setups A–D.

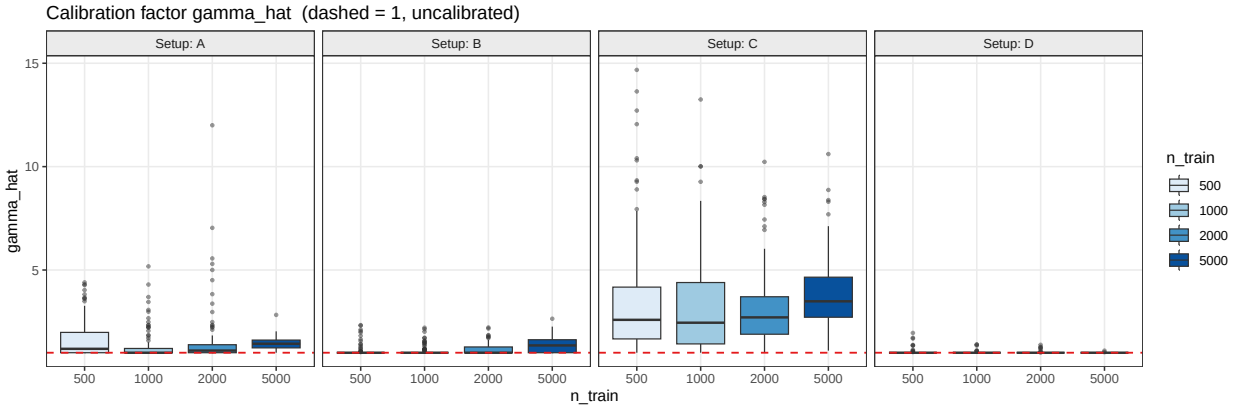

Figure 1.5: Calibration factor $\hat{\gamma}$ across training sample sizes and setups A–D. Dashed line at $\hat{\gamma} = 1$ (uncalibrated baseline).

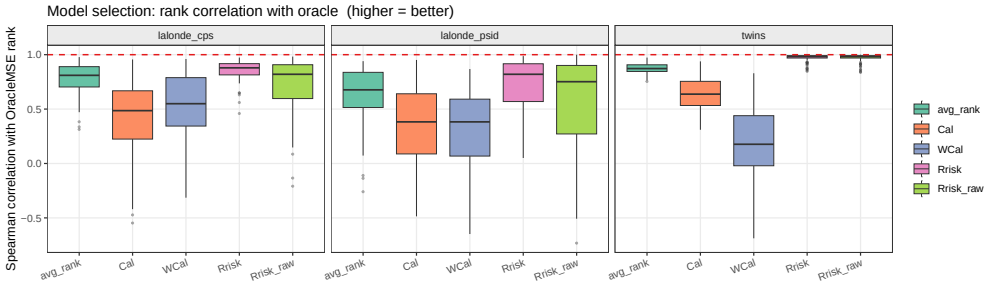

Figure 1.6: Spearman rank correlation between each screening metric and the oracle MSE ranking on the three RealCause semi-synthetic DGPs ($n_{\text{train}} = 1000$, 100 replications). Higher is better.

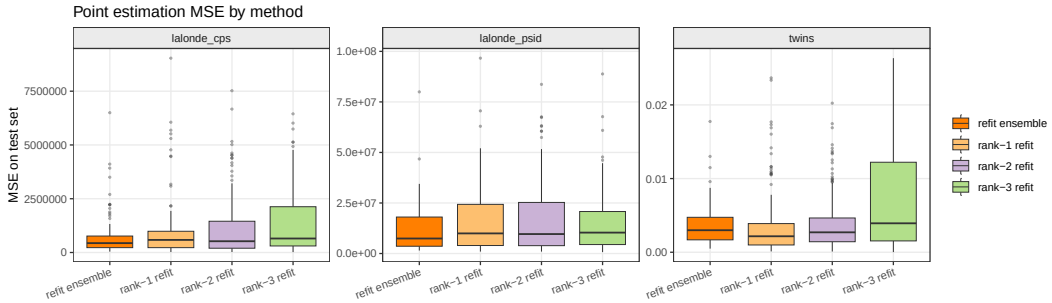

Figure 1.7: Test-set MSE of point estimators on the three RealCause DGPs.

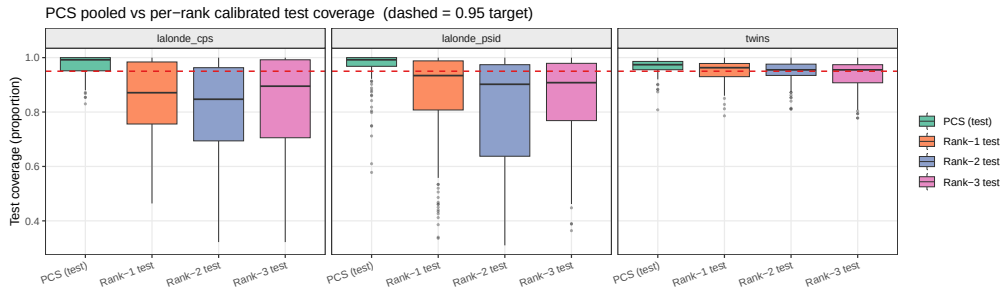

Figure 1.8: Test-set coverage of the PCS pooled and per-rank calibrated intervals on the three RealCause DGPs. Dashed line at the nominal 95% level.

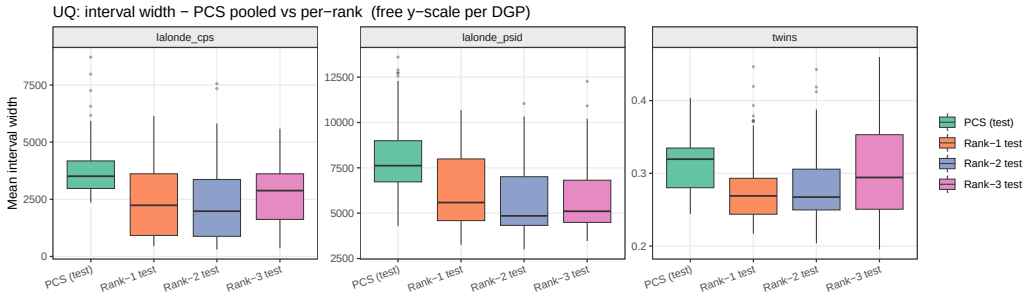

Figure 1.9: Mean interval width of the PCS pooled and per-rank calibrated intervals.

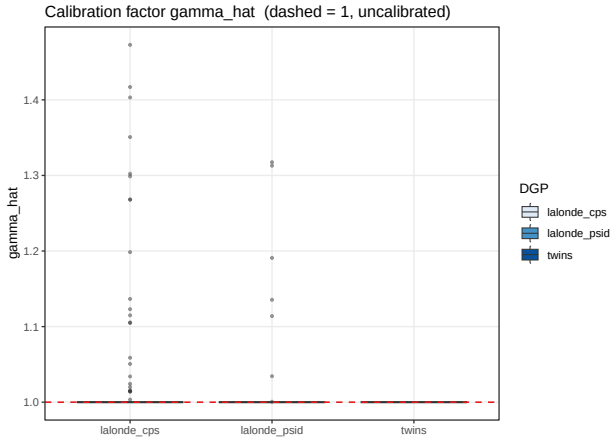

Figure 1.10: Calibration factor $\hat{\gamma}$ on the three RealCause DGPs. Dashed line at $\hat{\gamma} = 1$.

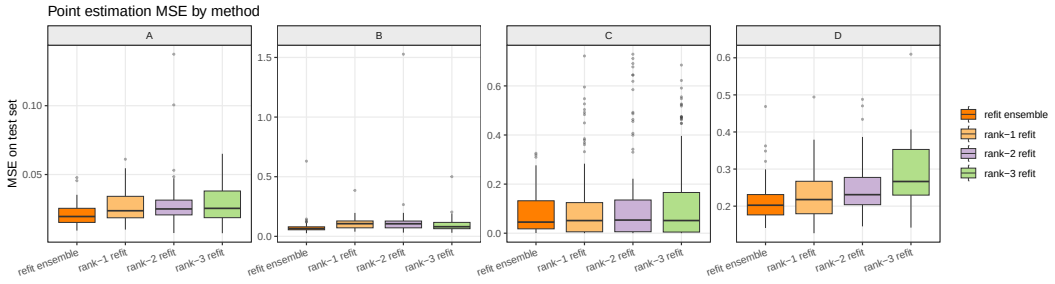

Figure 1.11: Test-set MSE of point estimators on observational-study DGPs.

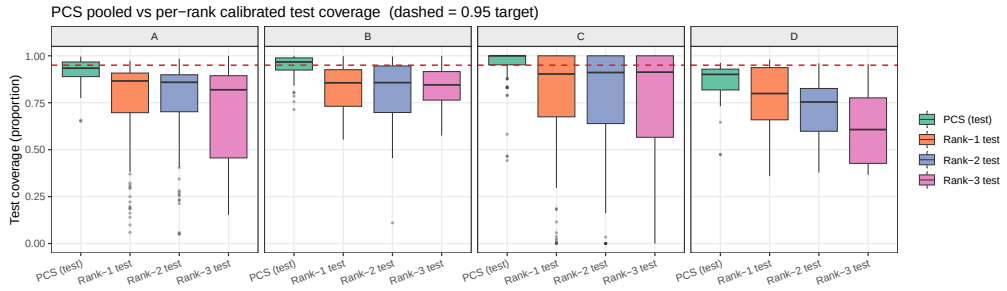

Figure 1.12: Test-set coverage of the PCS pooled and per-rank calibrated intervals on observational-study DGPs. Dashed line at the nominal 95% level.

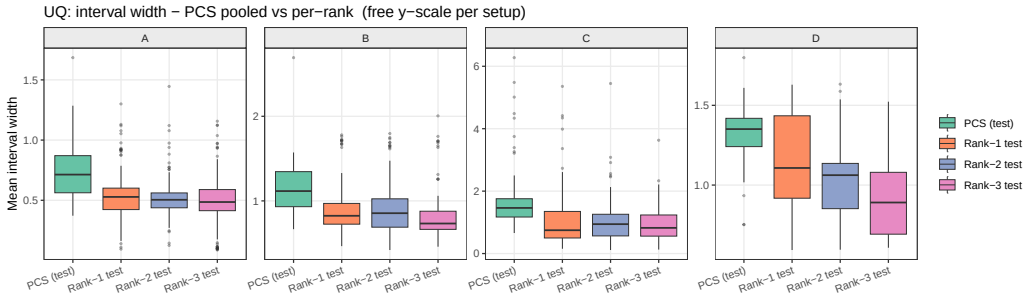

Figure 1.13: Mean interval width of the PCS pooled and per-rank calibrated intervals on observational-study DGPs.

Chapter 2

Unifying L_2 sensitivity analyses for regression and weighting estimators

2.1 Introduction

Observational studies are widely used to estimate causal effects, but their validity relies on assumptions that are inherently untestable. In particular, the assumption of no unobserved confounding is often implausible in practice. Sensitivity analysis addresses this concern by quantifying how violations of this assumption could affect causal conclusions. A large body of work has developed sensitivity analysis methods tailored to specific estimation strategies. In regression settings, omitted-variable approaches parameterize the strength of an unobserved confounder through partial R^2 measures [e.g., Cinelli and Hazlett [16]]. In weighting-based approaches, sensitivity is often characterized through perturbations of propensity scores or weights, including worst-case ratio bounds [23] or variance-based measures [39]. While these methods are well developed within their respective frameworks, they are typically designed for specific estimation strategies.

Recent work has highlighted a close connection between regression and weighting estimators: many regression estimators can be expressed as weighted averages with data-dependent implied weights [12]. This perspective suggests that sensitivity to unobserved confounding can be studied through perturbations of these weights, providing a common language across estimation strategies.

In this paper, we develop an L_2 sensitivity framework based on perturbations of implied weights. We consider an estimator of the form $\hat{\tau} = w^\top Y$, where w depends on observed covariates, and study deviations from an oracle weight vector w^* that would arise if unobserved confounders were observed. Our framework constrains the discrepancy $\|w^* - w\|_2$ and characterizes the resulting worst-case bias. We introduce two complementary formulations of this framework. The *restricted* model assumes that the oracle weights arise from augmenting the estimator with an omitted confounder, leading to interpretable sensitivity parameters and sharp bias bounds within this structured class. The *unrestricted* model instead treats w^*

as an arbitrary target weight vector in an L_2 neighborhood of w, providing valid worst-case bounds that are robust to model misspecification but less structurally interpretable.

Within this framework, we establish several connections to existing sensitivity analyses. For linear regression, the proposed sensitivity parameter coincides with the partial R^2 used in omitted-variable analyses, providing a geometric interpretation of classical results. More broadly, omitted-variable sensitivity analyses including Cinelli and Hazlett [16] for linear regression and Chernozhukov et al. [14] for more general causal machine learning settings, fall within the restricted formulation: the oracle quantity is defined by augmenting the observed specification with an omitted confounder U. For weighting estimators, it is in one-to-one correspondence with variance-based sensitivity parameters. The framework extends naturally to interacted linear regression, where standard sensitivity tools are less directly applicable, and to other estimators that admit implied weights. Our results provide a unified perspective on sensitivity analysis across regression and weighting estimators, clarify the role of structural assumptions in determining sharpness and interpretability, and offer a tractable approach to sensitivity analysis in settings with heterogeneous treatment effects or more flexible estimators.

2.2 Sensitivity analysis with L_2 sensitivity model

2.2.1 Setup and Notation

We consider n units randomly sampled from a population, indexed by $i = 1, \ldots, n$. For each unit, let X_i denote the observed covariates, $Z_i \in \{0, 1\}$ the treatment indicator, U_i an unobserved confounder, and $Y_i(1)$ and $Y_i(0)$ the potential outcomes under treatment and control, respectively. The observed data are therefore $\{(\mathbf{X}_i, Z_i, Y_i(Z_i))\}_{i=1}^n$, while U_i and the missing potential outcome remain unobserved. We study two commonly considered causal estimands: the average treatment effect $\tau_{ATE} = \mathbb{E}\{Y(1) - Y(0)\}$ and the average treatment effect on the treated $\tau_{ATT} = \mathbb{E}\{Y(1) - Y(0) \mid Z = 1\}$. In observational studies, where treatment is not randomly assigned, additional assumptions are required for identification. Three commonly used assumptions are the Stable Unit Treatment Value Assumption (SUTVA), overlap, and conditional ignorability.

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA))**.** *For each unit i, the observed outcome satisfies $Y_i = Y_i(Z_i)$, and there is no interference between units.*

Assumption 2 (Overlap)**.** *For all covariate values x in the support of X_i,*

$$0 < \Pr(Z_i = 1 \mid X_i = x) < 1.$$

Assumption 3 (Conditional ignorability)**.** *The potential outcomes are independent of the treatment assignment conditional on observed covariates:*

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i \mid X_i.$$

Under the above assumptions, researchers have developed various estimation strategies to estimate the causal estimands. Regression-based estimators are among the most widely used tools for causal inference. Common examples include linear regression, interacted linear regression, as well as regularized regression adjustments such as lasso and ridge-type methods [5, 46]. For example, one simply fits a linear regression model $Y_i \sim X_i + Z_i$ and uses the coefficient of Z_i as an estimator of ATE. However, simple linear regression is often insufficient because it imposes a constant treatment effect and, even in randomized experiments, can exhibit finite-sample bias when treatment effects are heterogeneous. Interacted linear regression relaxes this restriction by allowing treatment effects to vary with observed covariates, whereas more complex regression procedures may be preferable in settings that require regularization, weighting, or other forms of modeling flexibility.

Another commonly used estimating strategy is weighting. The idea is to weight the units by the inverse of the propensity score (or other sorts of balancing weights), and then use the weighted average as an estimator of the causal estimand, i.e., $\hat{\tau} = \sum_i w_i Y_i$. Recent work by Chattopadhyay and Zubizarreta [12] has connected the two strategies by showing that regression estimators can be viewed as weighting estimators, and the implied weights for regression estimators can be expressed as a linear function of the observed covariates and the treatment indicator. For example, in the simple linear regression of Y on $(1, Z, X)$, the OLS coefficient on Z can be written as $\hat{\tau} = w^\top Y$, where the signed implied weights are

$$w_i = \begin{cases} \dfrac{1}{n_t} + \dfrac{n}{n_c}(X_i - \bar{X}_t)^\top (S_t + S_c)^{-1}(\bar{X} - \bar{X}_t), & Z_i = 1, \\ -\dfrac{1}{n_c} - \dfrac{n}{n_t}(X_i - \bar{X}_c)^\top (S_t + S_c)^{-1}(\bar{X} - \bar{X}_c), & Z_i = 0, \end{cases}$$

with $S_t = \sum_{i:Z_i=1}(X_i - \bar{X}_t)(X_i - \bar{X}_t)^\top$ and $S_c = \sum_{i:Z_i=0}(X_i - \bar{X}_c)(X_i - \bar{X}_c)^\top$. They also showed that the linear regression weights satisfy basic balance properties, and that examining individual weights can serve as a regression diagnostic tool.

However, without randomization, the conditional ignorability assumption (Assumption 3) is often too strong if we believe there are unmeasured confounders, which is almost always the case in observational studies. To address this, sensitivity analysis relaxes the conditional ignorability assumption by allowing the potential outcomes to be dependent on the treatment assignment conditional on both observed covariates and an unobserved confounder.

Assumption 4 (Conditional ignorability with U)**.** *The potential outcomes are independent of the treatment assignment conditional on observed covariates and an unobserved confounder:*

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i \mid X_i, U_i.$$

Assumption 4 is more realistic and allows for the presence of unmeasured confounders, but because U_i is unobserved, the estimands of interest remains unidentifiable, and only using X_i will lead to biased estimates. Sensitivity analysis addresses this problem through partial identification: rather than assuming away the effect of U_i, one posits a sensitivity model

that constrains the magnitude of the distortion induced by omitting U_i, and then derives the resulting worst-case bias or identified set under a prespecified sensitivity parameter.

For linear regression, Cinelli and Hazlett [16] proposed a sensitivity analysis framework based on a partial R^2 parameterization of the omitted confounder. However, comparable sensitivity tools are not directly applicable to interacted linear regression, and existing approaches do not naturally extend to the broader class of regression estimators characterized through implied weights. Chernozhukov et al. [14] extends the omitted variable style and the partial R^2 parameterization to the context of causal machine learning. For weighting estimators, Dorn, Guo, and Kallus [23] proposed a sensitivity framework that bounds the worst-case ratio between oracle and estimated weights, w^*/w, corresponding to an L_∞ sensitivity model. In contrast, Huang and Pimentel [39] proposed a variance-based sensitivity model that can be interpreted as imposing an L_2 constraint on the deviation between oracle and estimated weights.

We propose a general L_2 sensitivity framework that applies to both regression and weighting estimators. The framework yields a simple and interpretable sensitivity analysis for interacted linear regression with sharp bias bounds, and naturally extends to more general regression estimators. It also bridges the regression and weighting sensitivity literatures, enabling a unified perspective and facilitating comparisons across estimation strategies

Throughout the main text, we condition on the realized sample and study sensitivity through perturbations of the realized implied weights. A population analogue is obtained by replacing sample implied-weight vectors with their corresponding population representers and Euclidean norms with $L_2(P)$ norms. Under this formulation, the sensitivity parameter will measure the distributional discrepancy between the oracle and observed representers, which aligns with the variance-based sensitivity model in Huang and Pimentel [39], and the finite-sample results are plug-in counterparts of the corresponding population bias bounds.

2.2.2 L_2 sensitivity model

We study sensitivity to unobserved confounding through perturbations of the weights, where weights are defined as w in the estimator $\hat{\tau} = w^\top Y$. Let w denote the weights from an estimator that conditions only on the observed covariates X, and let w^* denote the oracle weights that would arise if treatment ignorability held conditional on both X and an unobserved confounder U. The discrepancy induced by omitting U can therefore be represented as a perturbation of the weight vector. Our L_2 sensitivity framework constrains the magnitude of this perturbation through the L_2 norm and studies the resulting worst-case bias of the estimator.

Definition 1 (L_2 sensitivity model)**.** *For a given sensitivity parameter Γ, define*

$$\sigma(\Gamma) := \left\{ w^* : \|w^* - w\|_2^2 \leq f(\Gamma) \right\},$$

where $f(\cdot)$ is a known monotonically increasing function that maps the sensitivity parameter to the L_2 norm of the admissible perturbation.

This formulation allows the sensitivity parameter to be expressed through different parameterizations depending on the structure of the estimator. In the most general case, the model can be parameterized directly by a scalar c,

$$\sigma(c) := \left\{ w^* : \|w^* - w\|_2^2 \leq c \right\},$$

which imposes an L_2 bound on the deviation between the oracle and observed weights without requiring additional structural assumptions.

When additional structure on the weights is available, more interpretable parameterizations of $f(\cdot)$ can be introduced. In particular, for regression estimators the perturbation in the weights is induced by omitting a covariate U from the regression model. This additional structure allows the sensitivity parameter to be linked to interpretable quantities derived from the regression specification, as we illustrate in the following sections.

L_2 sensitivity model for linear regressions

For linear regression estimators, the general L_2 sensitivity model can be refined by exploiting the structure of regression implied weights. We consider two versions of the model: a *restricted* model and an *unrestricted* model.

The restricted model assumes that the oracle weights arise from augmenting the regression specification with an omitted covariate U. Under this structure, the deviation between the observed weights w and the oracle weights w^* is induced by omitting U from the regression. This structure allows the sensitivity parameter to be expressed through a weight-based quantity

$$R_w^2 := 1 - \frac{\|w\|_2^2}{\|w^*\|_2^2},$$

which measures the fraction of variation in the oracle implied weights attributable to the omitted confounder. This parameterization is closely related to the partial R^2 sensitivity parameter introduced by Cinelli and Hazlett [16] and the variance-based sensitivity parameter introduced by Huang and Pimentel [39]. Formally, the restricted sensitivity model is defined as

$$\sigma_{res}(\Gamma) := \left\{ w^* : \|w^* - w\|_2^2 \leq f(\Gamma), \text{ and } \exists U \text{ such that } w(X,U) = w^* \right\}.$$

Under this restricted model, the parameterization in terms of R_w^2 implies

$$\|w^* - w\|_2^2 = f(R_w^2) = \frac{R_w^2}{1 - R_w^2} \|w\|_2^2,$$

which is a monotone transformation of the sensitivity parameter. Equivalently, for a given R_w^2, the restricted sensitivity model can be written as

$$\sigma_{res}(R_w^2) := \left\{ w^* : 1 - \frac{\|w\|_2^2}{\|w^*\|_2^2} \leq R_w^2, \text{ and } \exists U \text{ such that } w(X,U) = w^* \right\}.$$

The restricted model therefore links the L_2 perturbation in weights to a regression-based interpretation of omitted-variable bias. In this case, the oracle weights have a concrete meaning: they are the implied weights from the same regression estimator, but computed under a richer conditioning set (X, U) that would satisfy conditional ignorability. Thus w^* is "oracle" relative to the estimator class under consideration: it is the weight vector that would be obtained if the analyst observed the confounder and fit the same regression specification with U included.

This interpretation becomes more delicate once we move beyond the restricted model. If the regression specification is misspecified, or if unobserved confounding cannot be represented as simply augmenting the regression with an additional covariate U, then there may be no unique regression-implied oracle weight vector. In that case, it is more natural to interpret w^* as a latent target weight vector that would deliver the estimand under a benchmark "truth"—for example, a fully adjusted weighting rule or an idealized regression adjustment that is unavailable to the analyst. The unrestricted model deliberately leaves this mechanism unspecified and only requires that the discrepancy between w^* and w be small in L_2 norm.

Accordingly, the distinction between the two models is not only whether the perturbation is structured, but also what object is being treated as the target. Under the restricted model, w^* is a regression-implied oracle weight vector generated by adding an omitted confounder to the same estimator. Under the unrestricted model, w^* should be viewed more abstractly as any target weight vector that is consistent with the estimand of interest and lies within an L_2 neighborhood of the observed weights. The unrestricted model therefore accommodates both omitted-variable bias and departures from the assumed regression class, at the cost of giving up the tighter structural interpretation that underlies the R_w^2 parameterization and the sharp bounds below.

This distinction also helps locate existing sensitivity analyses within our framework. Omitted-variable analyses based on short and long regressions, including Cinelli and Hazlett [16] and the more general causal machine learning framework of Chernozhukov et al. [14], belong to the restricted class because the oracle quantity is obtained by augmenting the observed specification with an omitted confounder U. By contrast, marginal sensitivity models [93, 23], originating from inverse probability weighting formulations of causal effects, restrict the admissible oracle weights through bounds on treatment assignment probabilities. These constraints can be derived from a class of data-generating processes with unobserved confounding, but they do not require that the oracle weights arise from augmenting the same estimator with an omitted covariate U. In this sense, MSM-based sensitivity analyses constrain the feasible set of weights directly, rather than specifying a regression-based mechanism for generating w^*, and are therefore closer in spirit to our unrestricted formulation.

The distinction matters because it reflects two different notions of sensitivity: sensitivity to omitted-variable bias within a specified model (restricted), versus sensitivity to arbitrary deviations from that model (unrestricted). The restricted model yields interpretable sensitivity parameters and sharp bounds but relies on structural assumptions, whereas the unrestricted model relaxes these assumptions and provides valid worst-case guarantees at

the cost of less interpretability and more conservative conclusions.

For practitioners, this distinction clarifies how sensitivity conclusions depend on modeling assumptions: if the linear specification is viewed as credible, the restricted analysis offers interpretable benchmarks and sharper conclusions; if there are concerns about misspecification or broader forms of confounding, the unrestricted analysis provides a more conservative assessment of sensitivity. In practice, reporting both analyses can be informative, as their comparison helps understand whether sensitivity is primarily driven by omitted-variable bias within the assumed model or by broader model dependence.

Simple linear regression (LR) We begin with the simplest regression specification, the linear regression of Y on (Z, X). In this setting the coefficient of Z is a widely used estimator of the average treatment effect (ATE). Using the implied-weight representation of linear regression, this estimator can be written as $\hat{\tau} = w^\top Y$, where w denotes the signed regression weights determined by (Z, X). When an unobserved confounder U is omitted from the regression, the implied weights change from w to the oracle weights w^* that would arise if the regression were augmented with U. The bias of the estimator can therefore be expressed as

$$\hat{\tau}^* - \hat{\tau} = (w^* - w)^\top Y = (\Delta w)^\top Y,$$

The next theorem characterizes the resulting bias under both the restricted and unrestricted L_2 sensitivity models. The restricted bound is sharp within the class of oracle weights generated by augmenting the regression with an omitted confounder, whereas the unrestricted bound is a general worst-case bound over an L_2 neighborhood of admissible target weights.

Theorem 1. *Under the restricted L_2 sensitivity model $\sigma_{res}(R_w^2)$ for LR, the bias satisfies*

$$|\hat{\tau}^* - \hat{\tau}| \leq \|Y^{\perp\{\mathbf{1},X,Z\}}\|_2 \, \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}}.$$

Moreover, the bound is sharp within the restricted oracle class. If we allow model misspecification and consider the general L_2 sensitivity model $\sigma(c)$, the bias satisfies

$$|\hat{\tau}^* - \hat{\tau}| \leq \sqrt{c}\|Y\|_2.$$

The parameter R_w^2 provides an interpretable measure of sensitivity: it quantifies the fraction of variation in the oracle weights w^* that is attributable to the omitted confounder. Larger values of R_w^2 correspond to greater deviations between the oracle and observed weights, and therefore larger potential bias. Besides the magnitude of the discrepancy of the weights, the bias also depends on the residual variation in the outcome after removing the linear effects of $(1, X, Z)$. Intuitively, only the component of Y that cannot be explained by the observed regression model can be exploited by an omitted confounder to induce bias.

Here sharpness is defined following Dorn, Guo, and Kallus [23]. A bound is sharp if, for every admissible value of the sensitivity parameter, there exists a data-generating process

consistent with the sensitivity model under which the bound is attained. Under the restricted model, sharpness is meaningful because the admissible perturbations are those generated by augmenting the regression with an omitted confounder U. Under the unrestricted model, by contrast, the class of admissible perturbations is defined only through an L_2 neighborhood around w. If w^* is interpreted as an arbitrary target weight vector, the Cauchy–Schwarz bound is worst-case over that enlarged class and is valid by construction, but its sharpness depends on whether every perturbation in that neighborhood corresponds to a meaningful target weighting scheme for the estimand. For example, if w^* represents exact balancing weights, then not all perturbations within the neighborhood satisfy the balancing conditions, and the bound is therefore not always sharp. For this reason, we emphasize sharpness for the restricted model, whereas for the unrestricted model the main guarantee is validity and robustness to misspecification rather than attainability within a structured oracle class.

The next proposition shows that, in the LR setting, this weight-based sensitivity parameter coincides with the partial R^2 parameter used in the omitted-variable sensitivity analysis. This establishes a direct connection between the classical regression-based sensitivity framework and the weight-based L_2 formulation developed here.

Proposition 1. *(Connection with Cinelli and Hazlett [16]) In the simple linear regression (LR), the partial R^2 sensitivity parameter $R^2_{Z\sim U|X}$ is equivalent to the restricted L_2 sensitivity parameter R^2_w:*

$$R^2_{Z\sim U|X} = R^2_w$$

This equivalence provides a useful interpretation of the proposed framework. The L_2 formulation therefore offers an alternative geometric perspective on omitted-variable bias, while also extending naturally to regression estimators—such as interacted linear regression—for which the classical partial R^2 framework is less directly applicable.

Interacted linear regression (ILR) We next consider interacted linear regression, where the regression specification includes treatment-covariate interactions and allows heterogeneous treatment effects. In this setting, the implied weights depend on the target group, and the resulting estimator can target either the average treatment effect (ATE) or the average treatment effect on the treated (ATT). As in the LR case, the estimator admits a signed implied-weight representation $\hat{\tau} = w^\top Y$, where the weights w are determined by (Z, X) under the chosen regression specification. When an unobserved confounder U is omitted from the regression, the implied weights change from w to the oracle weights w^* that would arise if the regression were augmented with U.

Compared with the LR case, the structure of the weight perturbation is more complex because the balancing conditions are imposed separately within treatment groups. Consequently, the bias depends on the residual variation of the outcome within the treated and control samples after removing the linear effects of the covariates. We therefore derive separate bias bounds for the ATE and ATT estimands under the restricted and unrestricted

L_2 sensitivity models. The next theorem characterizes the resulting bias under both the restricted and unrestricted L_2 sensitivity models.

Theorem 2 (L_2 sensitivity bounds for ILR-ATE)**.** *Under the restricted L_2 sensitivity model $\sigma_{res}(R_w^2)$ for ILR-ATE, the bias satisfies*

$$|\hat{\tau}^* - \hat{\tau}| \leq \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}} \sqrt{\|Y_t^{\perp\{\mathbf{1}, X_t\}}\|_2^2 + \|Y_c^{\perp\{\mathbf{1}, X_c\}}\|_2^2}.$$

Moreover, the bound is sharp within the restricted oracle class. If we allow model misspecification and consider the general L_2 sensitivity model $\sigma(c)$, the bias satisfies

$$|\hat{\tau}^* - \hat{\tau}| \leq \sqrt{c}\, \|Y\|_2.$$

Theorem 3 (L_2 sensitivity bounds for ILR-ATT)**.** *Under the restricted L_2 sensitivity model $\sigma_{res}(R_w^2)$ for ILR-ATT, the bias satisfies*

$$|\hat{\tau}^* - \hat{\tau}| \leq \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}}\, \|Y_c^{\perp\{\mathbf{1}, X_c\}}\|_2.$$

Moreover, the bound is sharp within the restricted oracle class. If we allow model misspecification and consider the general L_2 sensitivity model $\sigma(c)$, the bias satisfies

$$|\hat{\tau}^* - \hat{\tau}| \leq \sqrt{c} \|Y_c\|_2.$$

The bias bounds in Theorems 2 and 3 highlight both similarities and differences with the LR case. Under the restricted sensitivity model, the structure of the bound differs because interacted linear regression imposes balancing conditions separately within treatment groups. As a result, the worst-case bias depends on residual outcome variation within groups after removing the linear effects of the covariates. For the ATE estimand, both treated and control residual variation contribute through the term $\sqrt{\|Y_t^{\perp\{\mathbf{1}, X_t\}}\|_2^2 + \|Y_c^{\perp\{\mathbf{1}, X_c\}}\|_2^2}$, whereas for the ATT estimand only the control residual variation appears because treated weights are fixed by construction.

Under the unrestricted L_2 sensitivity model, the bounds are analogous to those in LR. Since the model only constrains the L_2 magnitude of the weight perturbation, the worst-case bias reduces to a Cauchy-Schwarz bound involving $\|Y\|_2$ (or $\|Y_c\|_2$ for ATT). Thus, the differences between LR and ILR arise primarily from the structural restrictions of the regression model rather than from the L_2 sensitivity framework itself.

Although Cinelli and Hazlett [16] focuses on LR, the `sensemakr` documentation [15] notes that sensitivity analysis for interacted linear regression can be implemented using the Hirano-Imbens reparameterization, which centers covariates and expresses treatment interactions as additional regressors. In this representation, partial R^2 parameters again characterize the strength of an omitted confounder. The next result shows that the weight-based sensitivity parameter in our framework coincides with this grouped partial R^2 characterization.

Proposition 2. *(Connection with Cinelli and Hazlett [16] and Cinelli, Ferwerda, and Hazlett [15]) For ILR(ATE) setting, with the Hirano-Imbens reparameterization, we have*

$$R_w^2 = R_{Z\sim G|W}^2$$

where $W = [1, X^{centered}, Z \cdot X^{centered}]$ *and* $G = [U^{centered}, Z \cdot U^{centered}]$.

This result links the weight-based sensitivity parameter to the partial R^2 formulation used in regression-based sensitivity analysis. While Cinelli and Hazlett [16] and `sensemakr` focus on the ATE, the implied-weight representation clarifies the target estimand under heterogeneous treatment effects and we show that the same L_2 sensitivity model applies across ATE and ATT, as well as LR and ILR, with differences determined by the implied weights.

In the ILR-ATT setting, conditioning on the observed data, the variance-based sensitivity parameter in Huang and Pimentel [39] and the restricted L_2 sensitivity parameter R_w^2 are in one-to-one correspondence. Specifically, we have the following proposition.

Proposition 3. *(Connection with Huang and Pimentel [39]) Let* w_c *and* w_c^* *denote the vectors of control weights under the observed and oracle models, respectively, written in the unsigned form used by Huang and Pimentel [39]. Then*

$$R_{\mathrm{VBM}}^2 := 1 - \frac{\mathrm{Var}(w_c)}{\mathrm{Var}(w_c^*)} = 1 - \frac{\|w_c\|_2^2 - 1/n_c}{\|w_c^*\|_2^2 - 1/n_c},$$

whereas our parameter satisfies

$$R_w^2 = 1 - \frac{1/n_t + \|w_c\|_2^2}{1/n_t + \|w_c^*\|_2^2}.$$

Equivalently,

$$\frac{R_w^2}{1 - R_w^2} = \frac{\|w_c\|_2^2 - 1/n_c}{1/n_t + \|w_c\|_2^2} \cdot \frac{R_{\mathrm{VBM}}^2}{1 - R_{\mathrm{VBM}}^2}.$$

Thus, for fixed observed data, R_{VBM}^2 *and* R_w^2 *differ only by a known monotone transformation induced by the different definitions of ATT weights.*

The connection follows because, under ATT, the treated weights are fixed at $1/n_t$ and only the control weights are perturbed by omitting U. In the notation of Huang and Pimentel [39], the control weights sum to one, so their sample variance satisfies $\mathrm{Var}(w_c) = n_c^{-1}(\|w_c\|_2^2 - 1/n_c)$, and likewise for w_c^*. Our signed implied weights differ only by attaching the fixed treated component and a negative sign to the control component, so the two sensitivity parameters encode the same strength of perturbation up to the deterministic rescaling above.

2.3 Illustration with data examples and discussion

2.3.1 Data Descriptions

We illustrate the proposed L_2 sensitivity framework using two widely studied observational datasets: an analysis of fish consumption and blood mercury levels using the National Health and Nutrition Examination Survey and an analysis of exposure to violence and attitudes toward peace using a 2009 survey of Darfurian refugees in eastern Chad. In each application, we estimate treatment effects using linear regression (LR) without treatment-covariate interactions and interacted linear regression (ILR), which allows for heterogeneous treatment effects across covariate strata. For ILR, we consider both the average treatment effect (ATE) and the average treatment effect on the treated (ATT). This design evaluates the different estimands and specifications within a common sensitivity framework defined by the global parameter R^2_w.

NHANES: Fish consumption and blood mercury

We re-analyze data from the 2013-2014 National Health and Nutrition Examination Survey (NHANES) [57]. Following previous studies [102, 90, 39], we define high fish consumption as more than 12 servings of fish or shellfish in the preceding month and low consumption as 0 or 1 servings, restricting the analysis to these two groups. The treatment is an indicator for high consumption, and the outcome is log blood mercury levels. Covariates include demographic and socioeconomic characteristics (age, gender, race, education, income, and smoking history).

Under LR-ATE, the estimated effect is 1.98 (95% CI: [1.80, 2.19]). Allowing for heterogeneous treatment effects yields an ILR-ATE estimate of 1.79 (95% CI: [1.58, 2.03]) and an ILR-ATT estimate of 2.07 (95% CI: [1.86, 2.28]). All specifications produce positive and statistically significant effects.

Darfur: Exposure to violence and peace attitudes

Our second application uses survey data from Darfurian refugees residing in eastern Chad to estimate the effect of direct exposure to violence on pro-peace attitudes (see [35]; also used as an empirical illustration in [16]). The treatment indicator equals one if a respondent reports having been physically injured during attacks on their village in Darfur during the 2003-2004 conflict, and the primary outcome is a continuous index constructed via factor analysis of multiple survey items measuring willingness to support peace and reconciliation. Covariates include demographic characteristics, livelihood status prior to displacement, household characteristics, and measures of prior political participation.

Under LR-ATE, the estimated effect of direct harm on pro-peace attitudes is 0.049 (95% CI: [0.014, 0.087]). The ILR-ATE estimate is 0.048 (95% CI: [0.013, 0.086]), and the ILR-ATT estimate is 0.054 (95% CI: [0.018, 0.094]). Across specifications, the estimated effect of exposure to violence on pro-peace attitudes is positive and statistically significant.

2.3.2 Sensitivity analysis

For each dataset, estimator, and estimand, we conduct L_2 sensitivity analysis by varying the global parameter R_w^2 (restricted model), computing the sharp worst-case bias bounds derived in Section 2, and constructing bootstrap confidence intervals for bias-adjusted estimates. We report two robustness thresholds: the *threshold (point)*, defined as the smallest value of R_w^2 for which the worst-case bias-adjusted point estimate reaches zero, and the *threshold (CI)*, defined as the smallest value of R_w^2 for which the lower bound of the bias-adjusted bootstrap confidence interval crosses zero. Because the confidence interval accounts for sampling uncertainty in addition to worst-case bias, the threshold (CI) is weakly smaller than the threshold (point).

To aid interpretation, we benchmark R_w^2 by omitting each observed covariate in turn and calculating the induced change in implied weights. This provides a scale for assessing the strength of a hypothetical unobserved confounder relative to measured predictors of the treatment assignment.

We also conduct unrestricted sensitivity analysis with the NHANES data (Figure 2.7). In this case, the parameter $c = \|\Delta w\|_2^2$ does not admit the same structural interpretation as arising from an omitted confounder added to the regression. Instead, it measures the magnitude of perturbations in the implied weights directly. To provide a scale for c, we consider a benchmark obtained by omitting each observed covariate in turn and computing the resulting change in weights, $\|w^{(-j)} - w\|_2^2$. This is different from the formal benchmarking strategy [16] in the restricted model, but may still provide useful insight.

Relative to the restricted model, the unrestricted sensitivity analysis delivers a worst-case bound over all perturbations of a given magnitude. For a fixed value of $\|\Delta w\|_2$, the resulting interval is therefore always weakly wider, reflecting the absence of the structural restrictions that tie w^* to an augmented regression in the restricted model. In the NHANES application, this yields visibly more conservative intervals under the unrestricted analysis.

NHANES

For LR-ATE (Figure 2.1), the threshold (point) is 0.318 and the threshold (CI) is 0.28. This implies that 31.8% of the variance in the oracle implied weights would need to be attributable to an omitted confounder for the positive estimate to be reversed, and 28% of the variance in the oracle implied weights would need to be attributable to an omitted confounder for the confidence interval to include zero. In contrast, the largest benchmarked covariate is *education*, which induces $R_w^2 = 0.043$. Thus, even the more conservative threshold (CI) is more than six times the largest observed benchmark, indicating substantial robustness of the LR estimate. For ILR-ATE (Figure 2.2), the threshold (point) is 0.205 and the threshold (CI) is 0.15. The largest benchmarked covariate is *race*, with $R_w^2 = 0.151$. Here the two thresholds lead to a more nuanced conclusion: the threshold (point) remains above the strongest observed benchmark, but the threshold (CI) is essentially at benchmark level. Thus, relative to LR-ATE, inference under ILR-ATE is materially more sensitive to confounding of

roughly observed strength. For ILR-ATT (Figure 2.3), the threshold (point) is 0.402 and the threshold (CI) is 0.34, while the largest benchmarked covariate *education* induces only $R_w^2 = 0.053$. This again suggests substantial robustness.

We also report unrestricted sensitivity results for NHANES LR-ATE. Using $c = \|\Delta w\|_2^2$ as the sensitivity parameter, the unrestricted threshold (point) is 0.001321 and the unrestricted threshold (CI) is 0.001150. The largest informal benchmark, again obtained by omitting education, is $c = 0.000287$, so both unrestricted thresholds are roughly four times the largest observed perturbation. Figure 2.8 compares the restricted and unrestricted analyses on a common c scale. As predicted by the theory, for a fixed magnitude of weight perturbation the unrestricted bound is uniformly wider than the restricted bound, reflecting the loss of the structural restrictions imposed by the restricted model.

Overall, the NHANES application shows meaningful heterogeneity across specifications. LR-ATE appears robust under both the restricted and unrestricted analyses. ILR-ATE is notably more sensitive, with its threshold (CI) approximately equal to the strongest observed benchmark. ILR-ATT remains comparatively robust, with both thresholds far above observed benchmark magnitudes.

Darfur

In contrast, the Darfur estimates exhibit much greater sensitivity. For LR-ATE (Figure 2.4), the threshold (point) is 0.006 and the threshold (CI) is 0.001. The largest benchmarked covariate is herder status, which induces $R_w^2 = 0.011$. Thus, the threshold (point) is already below the strongest observed benchmark, and the threshold (CI) is an order of magnitude smaller still. The same pattern holds under ILR-ATE (Figure 2.5), where the threshold (point) is 0.005, the threshold (CI) is 0.001, and the largest benchmark is 0.011, again generated by herder status. For ILR-ATT (Figure 2.6), the threshold (point) is 0.012 and the threshold (CI) is 0.002, while the largest benchmark is 0.021. Across all three estimands, even the point-based threshold lies below the strongest observed benchmark, and the CI-based threshold is substantially smaller still. This reflects the fact that the baseline Darfur estimates are positive but only modestly separated from zero, so relatively small amounts of confounding are sufficient to eliminate either the point estimate or statistical significance.

2.3.3 Discussion

In both applications, LR and ILR deliver positive and statistically significant baseline estimates, but their robustness differs once thresholds are compared with observed benchmarks. In NHANES, LR-ATE remains robust under both the restricted and unrestricted analyses, and ILR-ATT is also robust: for both estimands, the point and CI thresholds remain comfortably above the strongest observed benchmark. ILR-ATE is more sensitive. Its point threshold still exceeds the largest benchmark, but its CI threshold is essentially equal to it, indicating that statistical significance could be overturned by confounding of similar strength to the most influential observed covariate. In Darfur, by contrast, all three estimands are

substantially more sensitive. For LR-ATE, ILR-ATE, and ILR-ATT, even the point threshold lies below the strongest observed benchmark, and the CI threshold is smaller still. Thus, in the Darfur application, confounding weaker than the most influential observed covariate would already suffice to eliminate statistical significance.

These two applications highlight that sensitivity is driven by both the strength of the signal and the geometry of the estimator, as reflected in the implied weights. The contrast between LR-ATE and ILR-ATE in NHANES is explained by their different balancing targets, which are visible in Figure 2.9. LR-ATE achieves pairwise balance: the weighted treated and control means are equal, but that common balance point need not coincide with the overall population mean. ILR-ATE instead reweights each treatment group separately to the population mean, which aligns with the target for the ATE under heterogeneous treatment effects. In NHANES, the sample is highly imbalanced in size (234 treated versus 873 controls), so the population mean is much closer to the control group than to the treated group for several covariates, especially race. As a result, ILR must move the treated group substantially farther than LR, leading to more dispersed implied weights. This is reflected in both the lower ILR-ATE robustness threshold and the larger race benchmark. In contrast, the Darfur application shows far more comparable raw covariate distributions between treatment groups (Figure 2.10), with LR's balance point is close to the population mean for most covariates. Consequently, LR-ATE and ILR-ATE in Darfur have similar implied-weight variances and sensitivity thresholds.

The same mechanism also clarifies how benchmark rankings relate to raw imbalance. In general, the rankings reflect the magnitude of raw imbalance: covariates with larger imbalance tend to generate larger benchmarks. However, this relationship is not exact, as correlations among covariates can redistribute their contribution once others are held fixed. In NHANES, income, education, and race are correlated, so omitting one of them may have a limited effect if the others already absorb part of its contribution to the treatment assignment. This helps explain why income has the largest raw imbalance but not the largest LR benchmark, while race is the dominant benchmark for ILR-ATE because it is especially important for moving the treated group toward the population target. By contrast, in Darfur the covariate distributions are more comparable across treatment groups and the observed covariates appear less collinear, so LR-ATE and ILR-ATE have similar implied-weight dispersion, similar thresholds, and similar benchmark rankings.

Overall, these examples illustrate two distinct sources of sensitivity. The first is the magnitude of the baseline estimate relative to its sampling uncertainty, as in Darfur. The second is the geometry of the estimator, encoded by the dispersion of the implied weights, as in NHANES where ILR-ATE requires more aggressive reweighting than LR-ATE.

For the comparison between the restricted and unrestricted models, we refer to Figure 2.8. The figure plots both analyses on a common $c = \|\Delta w\|_2^2$ scale, making the two analyses directly comparable. For any fixed magnitude of weight perturbation, the unrestricted bound is uniformly wider than the restricted bound, reflecting that it considers all admissible perturbations rather than those generated by augmenting the regression with an omitted confounder. Consequently, the unrestricted thresholds are larger when expressed on the same

scale, indicating that conclusions are more sensitive under the worst-case formulation. In the NHANES application, both analyses agree that the LR-ATE is robust to confounding. This comparison illustrates the practical tradeoff discussed in Section 2: the restricted model delivers sharper, more interpretable bounds under a structured omission mechanism, while the unrestricted model provides a more conservative assessment that is robust to broader forms of misspecification. In practice, reporting both can be informative and transparent, and practitioners can make their own judgment call incorporating domain knowledge.

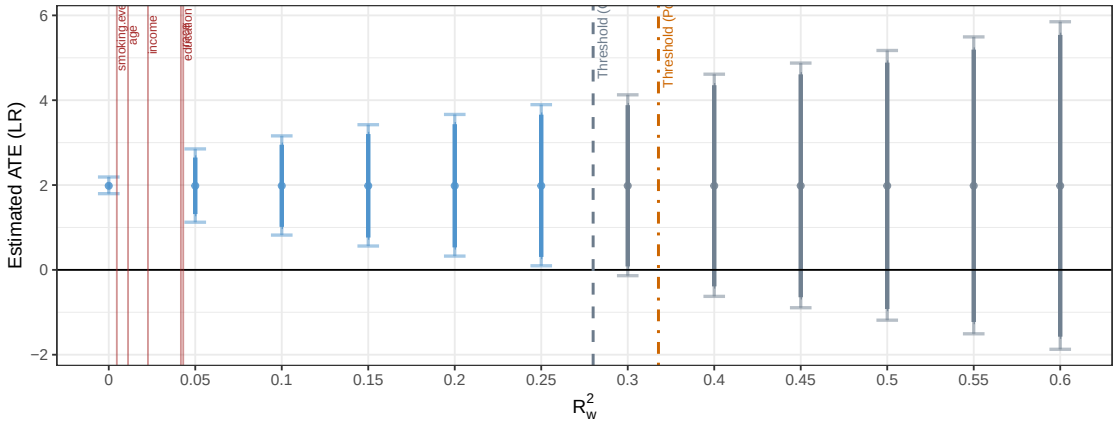

Figure 2.1: NHANES (LR ATE) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R^2_w varies. Yellow dashed line: threshold (point) = 0.318; gray dashed line: threshold (CI) = 0.28. Vertical markers indicate benchmarked covariates.

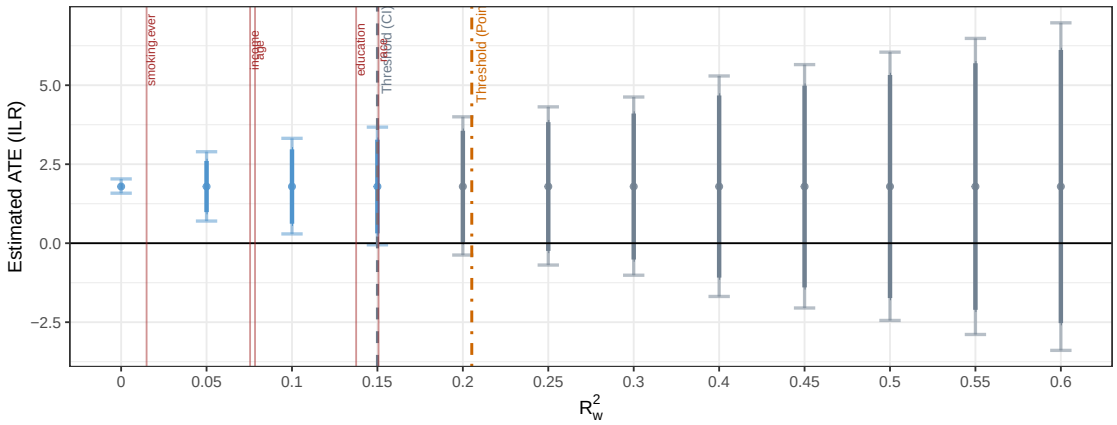

Figure 2.2: NHANES (ILR ATE) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R^2_w varies. Yellow dashed line: threshold (point) = 0.205; gray dashed line: threshold (CI) = 0.15. Vertical markers indicate benchmarked covariates.

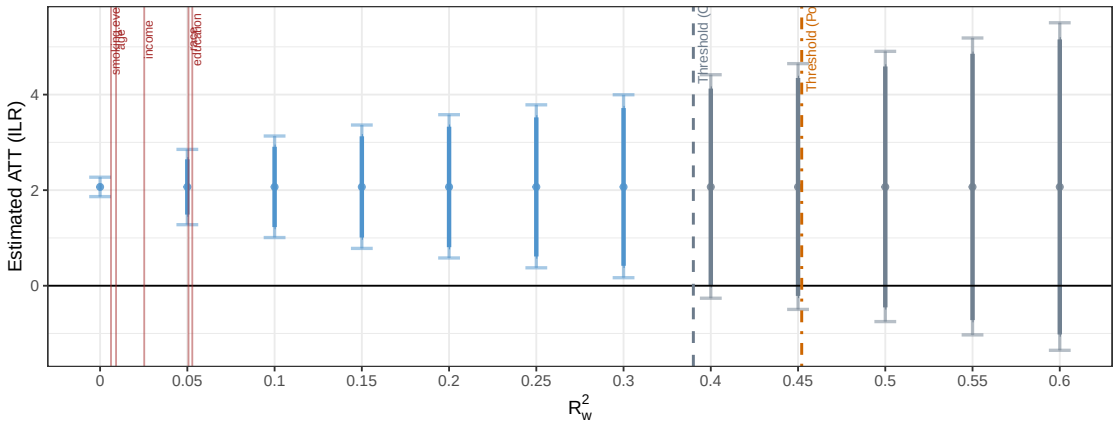

Figure 2.3: NHANES (ILR ATT) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R^2_w varies. Yellow dashed line: threshold (point) = 0.402; gray dashed line: threshold (CI) = 0.34. Vertical markers indicate benchmarked covariates.

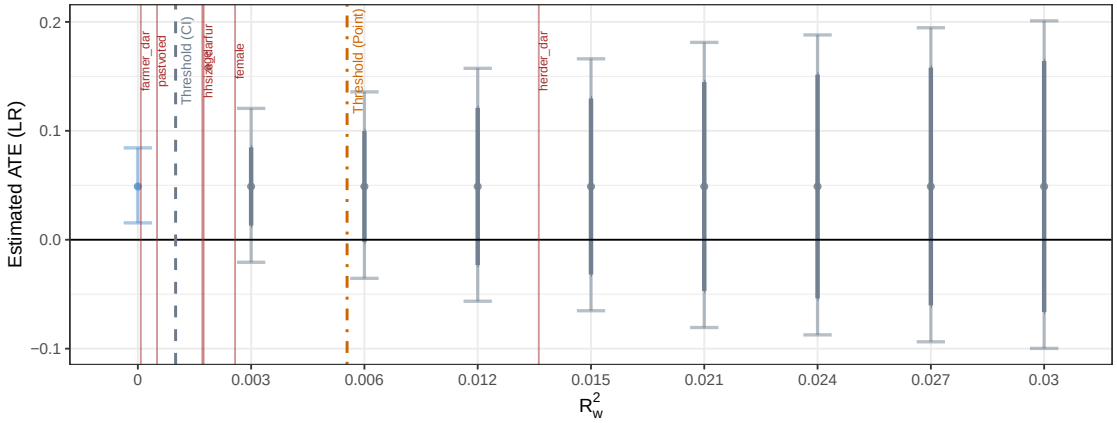

Figure 2.4: Darfur (LR ATE) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R^2_w varies. Yellow dashed line: threshold (point) = 0.006; gray dashed line: threshold (CI) = 0.001. Vertical markers indicate benchmarked covariates.

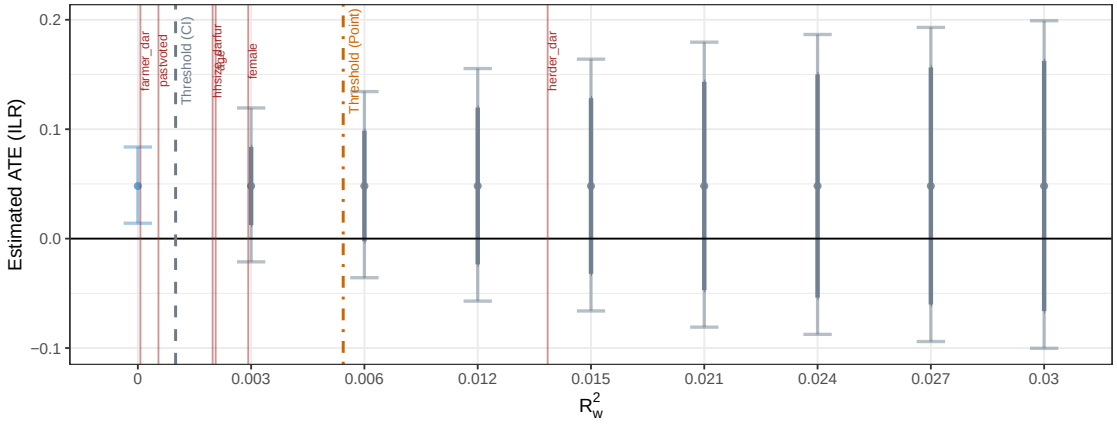

Figure 2.5: Darfur (ILR ATE) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R^2_w varies. Yellow dashed line: threshold (point) = 0.005; gray dashed line: threshold (CI) = 0.001. Vertical markers indicate benchmarked covariates.

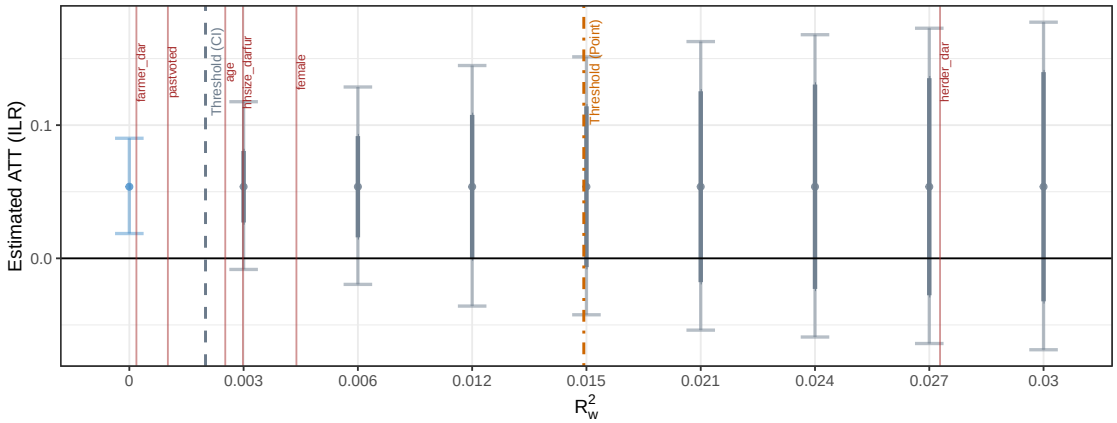

Figure 2.6: Darfur (ILR ATT) sensitivity analysis: bias-adjusted estimate and 95% bootstrap CI as R_w^2 varies. Yellow dashed line: threshold (point) = 0.012; gray dashed line: threshold (CI) = 0.002. Vertical markers indicate benchmarked covariates.

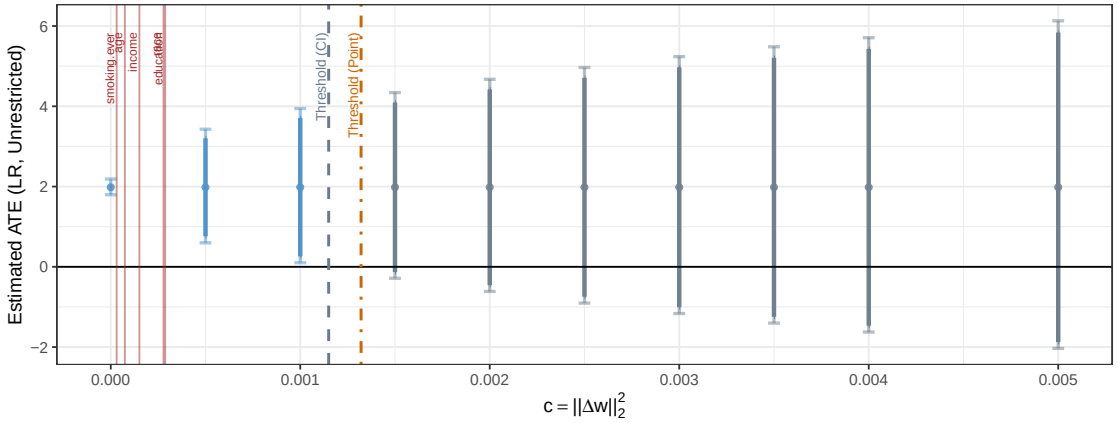

Figure 2.7: NHANES (LR ATE) sensitivity under the unrestricted L_2 model: bias-adjusted estimate and 95% bootstrap CI as the perturbation magnitude $c = \|w^* - w\|_2^2$ varies. Yellow dashed line: threshold (point) = 0.001321; gray dashed line: threshold (CI) = 0.001150. Vertical markers indicate benchmarked covariates.

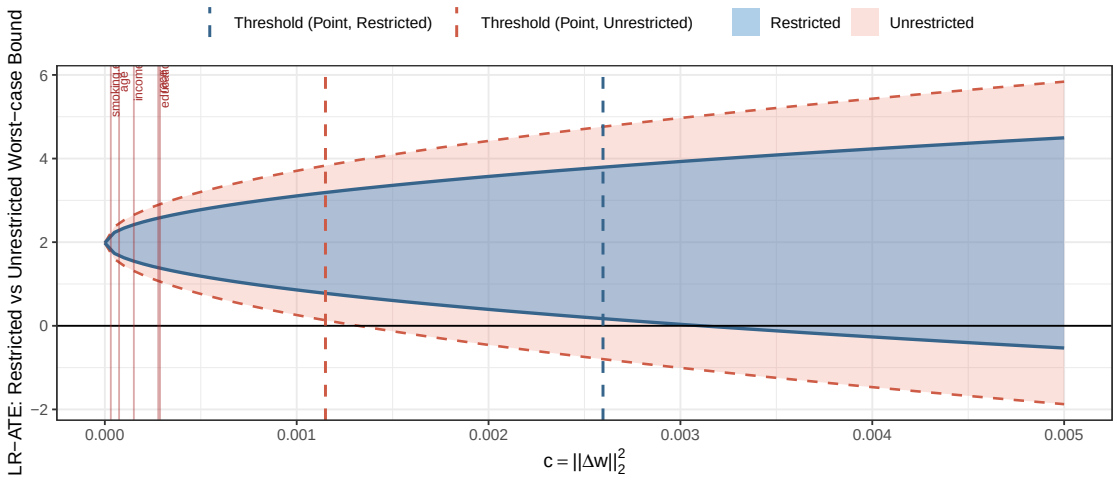

Figure 2.8: Restricted vs. unrestricted L_2 sensitivity for NHANES (LR ATE), plotted on a common $c = \|\Delta w\|_2^2$ scale; the unrestricted bound is uniformly wider at any fixed perturbation magnitude.

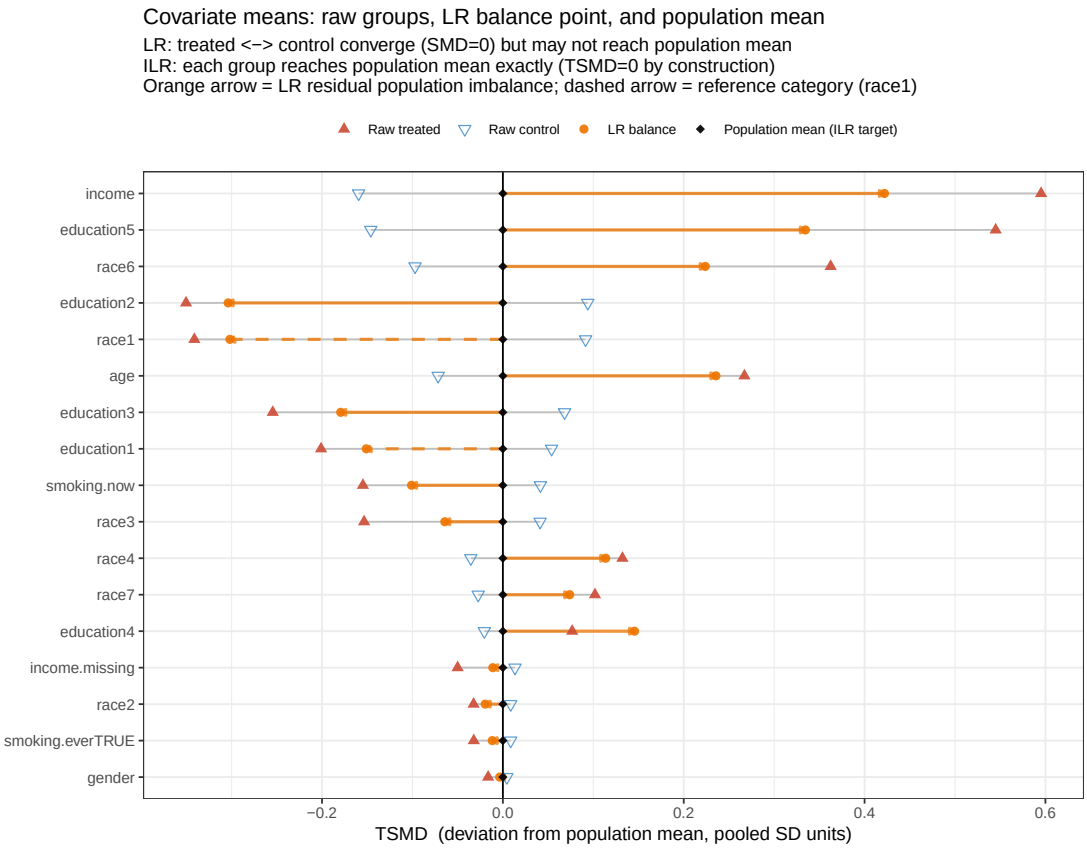

Figure 2.9: Covariate balance for NHANES: weighted treated, weighted control, and population means of each covariate under LR ATE, ILR ATE, and ILR ATT.

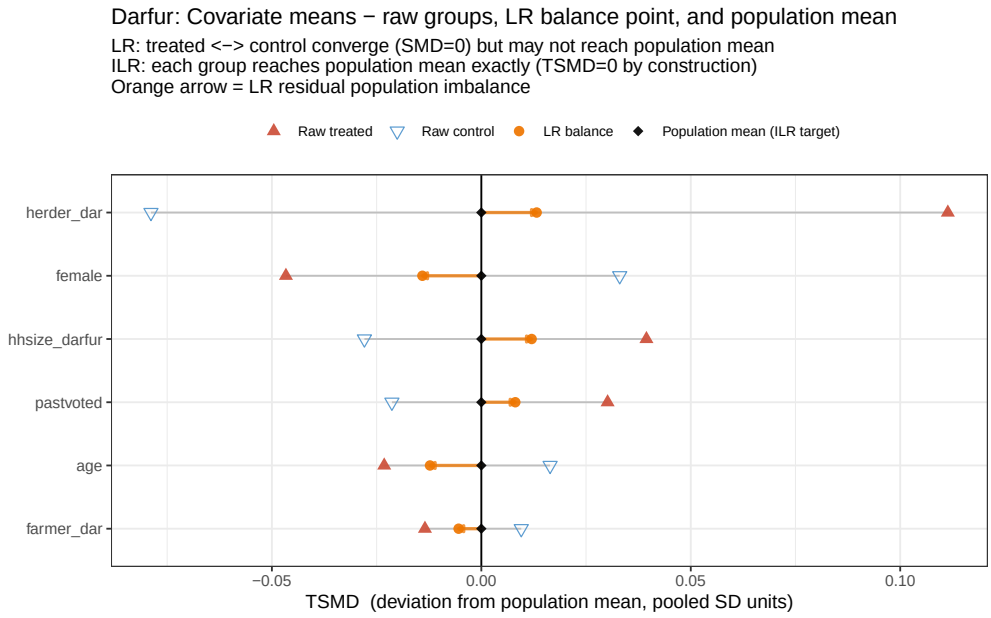

Figure 2.10: Covariate balance for Darfur: weighted treated, weighted control, and population means of each covariate under LR ATE, ILR ATE, and ILR ATT.

Chapter 3

Covariate-adaptive randomization inference in matched designs

3.1 Introduction

3.1.1 Re-evaluating a common model for inference

Randomized trials provide an effective means for measuring the effect of a treatment of interest relative to a control condition for at least two reasons. First, random allocation of treatment to units ensures that large differences in pre-treatment characteristics between the group of units selected for treatment and the group of units selected for control arise in large samples only with very small probability. Second, the known distribution of indicators of treatment across units in the study provides a basis for inference. By considering each unit's potential outcome values under treatment and under control as fixed latent values and repeatedly permuting treatment labels across study units (or using tests that rely on the large-sample behavior of such permutations) one may obtain inferences without making strong assumptions about the sampling procedure used to select the study units or the model for the outcome variable. This conceptual approach, dating back at least to Fisher [25], is often known as randomization inference.

In contrast, in observational studies of a binary effect concerns arise about whether units receiving treatment and units receiving control are otherwise comparable. If there is confounding, or systematic differences in variables (either observed or unobserved) that are predictive of the outcome of interest, the effect estimate from a simple group comparison will generally differ systematically from the effect that would have been measured in a randomized trial. To adjust for observed confounding variables, researchers may estimate an outcome model in the absence of treatment and compare units with similar expected outcomes under control, estimate a treatment model or propensity score and compare units with similar propensities to receive treatment, or some combination of the two. Matched observational studies adjust for observed confounding by grouping each treated unit with one or more similar control units and excluding controls not sufficiently similar to any treated

unit [92]. When matching is conducted without replacement of controls so that matched sets are disjoint, and with exact agreement on a propensity score so that units grouped together shared identical propensities for treatment, then each matched set is like a miniature randomized trial; conditional on one unit within the group receiving treatment, each is uniformly likely to have been the one selected. As such, methods of randomization inference are frequently applied to matched observational studies as though in a stratified randomized trial [88, 40, 87, 95]. In the language of Zhang and Zhao [100], the resulting procedures may be denoted "quasi-randomization tests."

Randomization inference in randomized trials depends on exact knowledge of the true randomization probabilities, and use of these methods in matched studies is motivated by an ideal setting in which the true propensity scores are known and matched exactly. In reality, however, propensity scores must be estimated, and except in cases where the measured variables are few and discrete they are never matched exactly. Optimal matching procedures such as those described in Rosenbaum [78], Zubizarreta [103], Austin and Stuart [7], and Pimentel et al. [68] use estimated propensity score differences as important inputs, so that when a match is created the researcher has information easily available about which matched sets are relatively closer or further from achieving the ideal uniform distribution for treatment assignment. Yet this information is not used when it comes time to do inference. By using uniform randomization inference, researchers implicitly assume a much simpler model and hope that these differences are all small enough not to create substantial lack of fit. While sensitivity analyses for unobserved confounding that are often conducted post hoc can in principle subsume discrepancies in observed variables too, these analyses are not typically presented in these terms, nor are observed differences typically used to calibrate the parameters for these sensitivity analyses.

We propose a new method for inference in matched observational studies, covariate-adaptive randomization inference, that explicitly uses estimated propensity score discrepancies to update permutation probabilities. This approach retains most of the advantages of uniform randomization inference – clear conceptual connections to a hypothetical randomized trial, ease of implementation, compatibility with interpretable methods of sensitivity analysis — while addressing potential for lack of fit even in settings where propensity score differences need not disappear in large samples. Furthermore, the lack of fit adjustment generally does not require users to alter the match itself in ways that reduce overall sample size. Although covariate-adaptive randomization inference can in theory fully resolve the confounding problem that matching itself seeks to address, we also demonstrate that matching followed by covariate-adaptive randomization offers precision and robustness benefits not enjoyed by covariate-adaptive randomization in an unmatched study.

3.1.2 Related work

Covariate-adaptive randomization inference builds on a quickly-growing literature that explores using estimated propensity scores to structure permutation tests. Rosenbaum [73] recommended permuting treatment assignment conditional on the sufficient statistic for a

propensity score fit, a closely related idea which works very well for settings with only one or two discrete covariates with a limited number of categories. In a graduate dissertation Baiocchi [8] briefly proposed permuting treatment assignments within matched pairs in a manner similar to that described above, although with a slightly different distribution based on a ratio of propensities on the probability scale rather than the odds scale. More recently, Branson and Bind [10] describe randomization tests with non-uniform treatment assignment probabilities for unmatched Bernoulli trials; although they focus on randomized trials they also suggest the use of the method for observational studies. The conditional permutation test of Berrett et al. [9] proposes using permutations of observed variables based on an estimated conditional distribution for the purposes of testing conditional independence. Shaikh and Toulis [86] also use permutations of observations based on estimated propensity scores to conduct causal inference in an observational study and obtain even stronger guarantees about large-sample performance by leveraging specific focus on a setting in which only one unit receives treatment. Resampling observations with non-uniform probabilities is also an important component of modern conformal inference for settings where data are not exchangeable [97], including observational studies of treatment effects [49]. Our contribution to this literature is to articulate the particular advantages of combining non-uniform permutation with matched designs. The structure of matched sets helps resolve many of the practical challenges that arise in conducting non-uniform permutation tests, such as the question of how to sample from the permutation distribution discussed at length by Berrett et al. [9]. In addition, covariate-adaptive randomization inference in concert with matching has important precision benefits relative to covariate-adaptive randomization in an unmatched study, as we will document in Section 3.6. In contrast to previous authors, we also integrate covariate-adaptive randomization inference into the sensitivity analysis framework of Rosenbaum [77], allowing us to conduct inference in the presence of unobserved confounding variables.

At a high level, covariate-adaptive randomization inference may also be compared to inverse probability weighting approaches to causal inference. Both methods rely on estimated propensity scores to adjust for observed confounding variables, but the two approaches use the estimated scores differently. In inverse weighting, estimators for a treatment effect are constructed by dividing each observed outcome by the estimated probability of its observed treatment (or some function thereof) before comparing average outcomes across groups. This corrects potential outcomes in expectation and produces unbiased estimates of average treatment effects [59], but can also lead to high-variance estimators since inverse weights can be large when propensity scores are near zero or one [72, 43]. In contrast, covariate-adaptive randomization inference uses estimated propensity scores to correct the null distribution of an arbitrary test statistic rather than to construct a specific test statistic Of course, inverse weighting estimators may also be used within the covariate-adaptive randomization inference framework, an idea we explore in Section A.1 of the online appendix.

Another difference between the inverse weighting framework and covariate-adaptive randomization inference is the type of null hypothesis typically tested. Weighting approaches usually test weak null hypotheses that only place restrictions on average potential outcomes,

remaining mostly agnostic to how treatment effects vary across individuals, while covariate-adaptive randomization inference as introduced here focuses on sharp null hypotheses that define a specific treatment effect for each individual [22]. While weak-null approaches account naturally for treatment effect heterogeneity, they also rely on asymptotic distributions of test statistics while a covariate-adaptive test of a sharp null does not (only requiring asymptotic convergence of the propensity score estimate to the true propensity score, a potentially weaker assumption when the propensity score is estimated in a separate dataset as discussed in Section 3.4 below). Weak null approaches generally require more restrictive conditions on test statistics than permutation tests of the sharp null, which are valid for arbitrary functions of treatment and outcome. Finally, an extensive sensitivity analysis framework has been developed for tests of the sharp null [77], which we adapt below in Section 3.5. Note also that tests designed for sharp null hypotheses can also be adapted to test weak nulls instead [11, 28], a point we discuss further in Section 3.8.

3.1.3 Outline

In what follows we develop and evaluate covariate-adaptive randomization inference. In Section 3.2 we introduce a formal framework and describe the shortcomings of uniform randomization inference in detail. In Section 3.3 we introduce covariate-adaptive randomization inference, giving procedures for hypothesis testing and for estimation, and confidence interval construction under a constant additive effect model. In Section 3.4 we bound the error introduced into the procedure by estimation of the propensity score. In Section 3.5, we generalize sensitivity analysis procedures grounded in uniform randomization inference to covariate-adaptive randomization inference to allow valid inference in the presence of unobserved confounding variables. In Section 3.6, we demonstrate finite sample performance of covariate-adaptive hypothesis tests and confidence intervals and compare to alternative strategies, including both matching approaches that permute subjects uniformly and unmatched approaches that use a form of covariate-adapative randomization inference. In Section 3.7 we demonstrate implications for practice by reanalysis of two observational datasets: one measuring genetic damage experienced by welders and one assessing the impact of right-heart catheterization on patient mortality. Finally, Section 3.8 highlights important questions and connections raised by this work and outlines opportunities for further research.

3.2 Formal framework and problem setup

3.2.1 Uniform randomization inference in matched designs

Consider a population of individuals each represented by a vector $(Y(1), Y(0), Z, X, U)$. Z is a binary indicator for membership in the treatment group, X is a vector of observed covariates, and U is an unobserved covariate. The values $Y(z)$ are potential outcomes under different treatment conditions as defined under the Neyman-Rubin causal model and the

stable unit treatment value assumption [82, 38], which specifies that an individual's outcome depends only on its own treatment status, rather than on the treatment status of other individuals, and that the only versions of treatment are 0 and 1. Only the information (Y, Z, X) is observed by the analyst, where Y represents the observed outcome $Y(Z)$. Let $\lambda(\mathbf{x}) = P(Z|X = x)$ be the conditional probability of treatment given observed covariates, or the propensity score, and let $\pi(x, u) = P(Z|X = x, U = u)$ be the true probability of treatment (which also depends on the unobserved U).

We assume individuals are first sampled independently from the population, and then formed into a matched design $\mathcal{M}$, consisting of K matched sets, on the basis of their treatment variables Z and covariates X alone. Each matched set, numbered $k = 1$ through K, contains exactly one treated individual and one or more control individuals. Individuals in set k are numbered $k1$ through kn_k, where n_k is the number of units in set k, in arbitrary order, so that we may refer to the treatment indicator of the ith unit in the kth matched set as Z_{ki}. We also define $n = \sum_{k=1}^{K} n_k$. Let vectors $\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{Y}, \mathbf{Z}, \mathbf{U} \in \mathbb{R}^n$ contain the observed univariate data ordered so that units in the same matched set are contiguous. Let matrix $\mathbf{X}$ contain all n vectors X_{ki} in its rows with an identical ordering; in addition, we abuse notation slightly to let $\lambda(\mathbf{X})$ represent the n-vector of true propensity scores. A matched design is said to be exact on a particular variable or quantity $\mathbf{v} \in \mathbb{R}^n$ if for any matched set k, $v_{ki} = v_{kj}$ for all $i, j \in \{1, \ldots, n_k\}$. Let $\mathcal{Z}_{\mathcal{M}}$ be the set of all treatment vectors $\mathbf{Z}'$ such that $\sum_{i=1}^{n_k} Z'_{ki} = 1$ for all matched sets k.

We now consider the the distribution of treatment assignments conditional on the match selected and the potential outcomes, i.e $P(\mathbf{Z} \mid \mathcal{F})$ where $\mathcal{F} = \{\mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\}$. Rosenbaum [77, §3] argues that this distribution is discrete uniform on $\mathcal{Z}_{\mathcal{M}}$ under two key assumptions: first, the absence of unobserved confounding, under which $\lambda(x) = \pi(x, u)$ for all u, and exact matching on covariates $\mathbf{X}$ (or more generally exact matching on true propensity scores $\lambda(X)$). The null distribution for a test statistic T can then be derived under the following sharp null hypothesis of no treatment effects for any individual in the sample:

$$\mathbf{Y}(1) = \mathbf{Y}(0). \tag{3.1}$$

In particular, the sharp null hypothesis of no effect guarantees that the actual observed outcome Y_{ki} for individual i in matched set k would still have been observed had Z_{ki} taken on a different value. Thus under the sharp null the test statistic $T(Z, Y)$ varies conditional on $\mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)$ only through the vector $\mathbf{Z}$, which is uniformly distributed over all permutations $\mathbf{Z}_{perm}$ of the observed element of $\mathbf{Z}$ within matched sets. The exact p-value for the test of the sharp null, where we reject for larger values of $T(Z, Y)$, can be computed as the proportion of values of $\mathbf{Z}_{perm}$ for which $T(Z_{perm}, Y)$ exceeds $T(Z, Y)$. In practice this quantity can be computed via repeated Monte Carlo draws from the permutation distribution or via a normal approximation to the permutation distribution. For example, Rosenbaum [77][§2] gives asymptotic distributions for a variety of rank statistics.

Estimates of causal effects and confidence intervals can also be developed from the null distribution associated with a sharp null hypothesis. Under a treatment effect model such

as the constant additive model, the randomization test can be inverted to produce Hodges-Lehmann point estimates and corresponding confidence intervals. Alternatively, when the test statistic itself is an estimator for an effect of interest (as in the case of the difference-in-means estimator), confidence intervals can be obtained under a normal approximation in large samples by using a variance estimate obtained from the null distribution and quantiles of the normal probability density function.

3.2.2 Propensity score discrepancies imperil the uniform treatment distribution

While the assumption of exact matching on a true propensity score yields convenient mathematical symmetry matching is never exact on all measured covariates in datasets of any substantial scale, nor is it practical to match exactly even on a univariate true propensity score, both because propensity scores must be estimated in practice and because they are often modeled as smooth functions of continuous variables that do not agree exactly for any two subjects. As such it is important to consider the potential for resulting lack of fit between the nominal uniform distribution of treatment used for inference and the true distribution for a given matched design. Hansen [34] addressed this question for the difference-in-means estimator of an additive treatment effect and found potential for slow-shrinking finite sample bias when the true propensity score is not matched exactly. Other authors have shown that bias in treatment effect estimation [84] and failure of Type I error control for the uniform randomization test [31] persist even in infinite samples in settings where all treated units are matched in pairs without replacement except in unusual situations such as populations where the probability of propensity scores exceeding 0.5 is zero or the true outcome model is known. Of course, even bigger problems may arise if unobserved confounding is also present, but methods of sensitivity analysis have been constructed with specific attention to this issue, as will be explored in greater detail in Section 3.5.1.

It has been argued that certain kinds of balance tests may be understood as tests for lack of fit between the actual distribution of treatment and the uniform model of inference used [34], so that if balance tests do not reveal problems then this problem can be ignored. This approach falls short of resolving the problem optimally for two reasons. First and most importantly, the theory underlying these balance tests relies on an asymptotic regime in which propensity score differences within matched sets approach zero as sample size increases, which in turn comes from a pattern of ever-larger concentrations of control units in a region arbitrarily close to any given treated unit. In many settings, this assumption is not reasonable. For example, Pimentel et al. [67] consider the common setting under which matches are conducted within natural blocks or groups of units, such as matching patients within hospitals or students within schools. When, as in these examples, the size of an individual block may reasonably be viewed as bounded and the most natural way to think about increasing sample size is by adding more blocks, there is no reason to expect propensity score differences within matched sets to shrink to zero, since the concentration of

matchable controls near a given treated unit is limited by the upper bound on the size of a block. More generally, Sävje [84] demonstrated that when propensity scores larger than 0.5 are present in the population with probability exceeding 0 and matching is conducted without replacement, then some matched sets will necessarily have propensity score discrepancies bounded away from zero. A second problem with the balance testing solution is that it does not fully articulate how to resolve problems with a matched design when the balance test fails. Common solutions such as using a tighter propensity score caliper lead to tradeoffs by reducing other aspects of match quality such as the proportion of treated units retained.

Another proposed strategy is the use of regression adjustment to remove slow-shrinking bias not addressed by close matching on a propensity score or on covariates [1, 31]. Under assumptions on the outcome model, this method removes the bias, and under sufficiently strong assumptions on the outcome model and the convergence of matched discrepancies to zero, a fast rate is achieved. However, the assumptions may not always be plausible, particularly the condition that the propensity score discrepancies shrink to zero as discussed by Sävje [84]. In addition, estimating an outcome model may be inconvenient if the outcome is multivariate, is related to observed covariates in a complex or poorly-understood manner, or is not yet measured at the time the match is conducted.

3.2.3 Lack of fit due to dependence between the true treatment vector and the match itself

All of the above discussion focuses on discrepancies between the uniform distribution of treatment given $\mathcal{F} = \{\mathcal{Z}_\mathcal{M}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\}$ and the actual distribution of treatment given these quantities, thinking of the match $\mathcal{M}$ as fixed over all possible $\mathbf{Z}$-values. However, if a model of independently-sampled subjects is assumed on the original data prior to matching then one might view the matched design, which is constructed with reference both to $\mathbf{X}$ and $\mathbf{Z}$, as a random function of treatment. In the special case of exact matching this issue need not arise, as the specific match chosen may remain conditionally independent of $\mathbf{Z}$ given event $\mathcal{Z}_\mathcal{M}$. However, when matching is not exact this guarantee no longer holds, and for some $\mathbf{Z} \in \mathcal{Z}_\mathcal{M}$ it may be the case that match $\mathcal{M}$ never would have been selected. For example, a treated unit with a high propensity score may be difficult to match and be paired with a control having an appreciably lower propensity score; however, if the treated unit had been a control instead, it would have been possible to form a better match with a treated unit having a similarly high propensity score value. Values of $\mathbf{Z}$ that would have produced a different match do not belong in the support of the distribution of treatment conditional on $\mathcal{M}$, but they appear with positive support in the other two distributions mentioned. The phenomenon of reduced support will be denoted Z-dependence (with reference to the dependence of the match chosen on the original treatment status vector).

This issue is also discussed by Pashley, Basse, and Miratrix [64], who note, "proper conditional analysis would need to take into account the matching algorithm." However, no such analysis has yet been proposed in the matching literature, and Pashley, Basse,

and Miratrix suggest that matching may perform well empirically even in the presence of Z-dependence. Z-dependence is not a central focus in what follows, and unless otherwise noted we will take the perspective that the matched sets $\mathcal{M}$ are fixed. However,in Section 3.6.4, we explore a "proper conditional analysis" that does account for the matching algorithm in a very small simulated dataset using brute-force methods, and find potential for empirically meaningful type I error violations. Z-dependence will also be helpful in explaining the empirical performance of uniform and covariate-adaptive randomization inference in certain larger simulation settings considered in Sections 3.6.2-3.6.3. We advocate and outline ideas for further progress in understanding Z-dependence in Section 3.8.

3.3 Adapting randomization inference to covariate discrepancies

3.3.1 True and estimated conditional distributions of treatment status

To adapt randomization inference to discrepancies in observed covariates, we begin by considering the case where no unobserved confounding is present and represent the true conditional distribution in terms of propensity scores. Since treatment indicator Z_{ki} is Bernoulli($\lambda(X_{ki})$), we have:

$$\begin{aligned}
&P\left\{Z_{ki} = 1 \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\right\} \\
&= \frac{P\left\{Z_{ki} = 1, Z_{k2} = 0, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k)\right\}}{\sum_{j=1}^{n_k} P\left\{Z_{k1} = 0, \ldots Z_{k(j-1)} = 0, Z_{kj} = 1, Z_{k(j+1)} = 0, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k)\right\}} \\
&= \frac{\lambda(X_{ki}) \prod_{j \neq i} [1 - \lambda(X_{kj})]}{\sum_{j=1}^{n_k} \lambda(X_{kj}) \prod_{\ell \neq j} [1 - \lambda(X_{k\ell})]} = \frac{\text{odds}\{\lambda(X_{ki})\}}{\sum_{j=1}^{n_k} \text{odds}\{\lambda(X_{kj})\}} = p_{ki}.
\end{aligned}$$

Because individuals are sampled independently, the joint conditional probability for a treatment vector $\mathbf{Z}$ is given by multiplying the appropriate p_{ki} terms together:

$$P\left\{\mathbf{Z}_{\mathcal{M}} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)\right\} = \begin{cases} \prod_{k=1}^{K} \prod_{i=1}^{n_k} p_{ki}^{z_{ki}} & \mathbf{z} \in \mathcal{Z}_{\mathcal{M}} \\ 0 & \mathbf{z} \notin \mathcal{Z}_{\mathcal{M}} \end{cases} \tag{3.2}$$

If the true propensity score $\lambda(\cdot)$ were known, the p_{ki}s could be calculated exactly. To conduct inference under the sharp null, the Monte Carlo strategy described in Section 3.2.1 could be used, except that instead of permuting treatment assignments within matched sets uniformly at random one would draw the identity of the treated unit in each set from an independent multinomial random variable with 1 trial and probabilities $p_{k1}, \ldots, p_{kn_k}$. Exact finite-sample confidence intervals could also be obtained by inverting the test, and the large majority of the benefits of the uniform randomization inference procedure could be retained despite the

reality of inexact matching on the propensity score. Note that when propensity scores are matched exactly, this procedure reduces to the uniform test.

Unfortunately, the propensity score is typically unknown, so the true conditional distribution of treatment is also unknown. However, it is straightforward to construct a plugin estimator for this distribution by substituting an estimate of the propensity score $\widehat{\lambda}(\cdot)$ for $\lambda(\cdot)$ in the formula for p_{ki}. We denote the process of conducting hypothesis tests and creating confidence intervals using this plugin distribution as covariate-adaptive randomization inference. In concrete terms, an investigator may conduct covariate-adaptive randomization inference via the Monte Carlo approach of Section 3.2.1 by calculating the estimated propensity odds for each subject in a matched set, by dividing each estimated odds by the sum of the odds in the matched set, by determining treatment via a single draw from a multinomial distribution with probabilities for each subject given by these normalized odds, and by conducting this process independently within all matched sets, repeating as necessary to generate a close approximation to the null distribution.

3.3.2 The difference-in-means estimator: large-sample distribution

While taking repeated Monte Carlo draws from the covariate-adaptive permutation distribution is straightforward, it is also instructive to construct a normal approximation using the mean and variance of the test statistic. Here we demonstrate this approach for the standard difference-in-means estimator, defined below (for brief discussion of an alternative estimator based on inverse propensity weighting, see Section A.1 of the online appendix).

$$T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) = \frac{1}{K}\sum_{k=1}^{K}\left\{\sum_{i=1}^{n_k} Z_{ki}Y_{ki} - \frac{1}{n_k-1}\sum_{i=1}^{n_k}(1-Z_{ki})Y_{ki}\right\} = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1}.$$

Under the sharp null hypothesis, the $\mathbf{Y}$ is fixed across all values for $\mathbf{Z}$ so the expectation and the variance are calculated only over the random variables $\mathbf{Z}$.

$$E(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}(1), \mathbf{Y}_{\mathcal{M}}(0)) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k p_{ki}-1}{n_k-1} \tag{3.3}$$

$$\begin{aligned} &Var(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}(1), \mathbf{Y}_{\mathcal{M}}(0)) \\ &\quad = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\frac{n_k}{n_k-1}\right)^2 Y_{ki}p_{ki}\left\{Y_{ki}(1-p_{ki}) - \sum_{j\neq i}^{n_k} Y_{kj}p_{kj}\right\} \end{aligned} \tag{3.4}$$

Note that these formulas get considerably simpler in the case of a matched pair design, in which $n_k = 2$ for all k. Here the inner sum $\sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1}$ may be rewritten as $(Z_{k1} - Z_{k2})(Y_{k1} - Y_{k2})$, so that the mean of the test statistic is the sample average (across matched pairs k) of $V_k D_k$ where $V_k = p_{k1} - p_{k2}$ and $D_k = Y_{k1} - Y_{k2}$, and the variance is the sample average of $K^{-1}(1 - V_k^2)D_k^2$.

While this mean and variance of the test statistic depend on the unknown true propensity score through the p_{ki} terms, they can be estimated by substituting the $\widehat{p}_{ki}$ obtained by substituting the estimated propensity score $\widehat{\lambda}$ for the true propensity score λ. Large-sample inference may be conducted by computing a normalized deviate of the difference-in-mean statistic using estimates of the mean and variance above, and comparing to the standard normal distribution. This relies on a finite-sample central limit theorem based on a sequence of infinitely-expanding finite samples [52] for which the mean and variance of the test statistic converge to stable limits. The most natural asymptotic regime places some upper bound on the size of a matched set n_k across all these samples while K grows towards infinity, since returns to matching large numbers of controls to a single treated unit are quickly diminishing [33]. The simplest available central limit theorem is thus in the inner sums $\sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1}$, which are independent random variables with non-identical distributions; when n_k is uniformly bounded and a Lindeberg condition holds on the potential outcomes $Y(0)$, they follow from results in Liu and Yang [55]. For more discussion of asymptotic regimes in matching with restricted or fixed matching ratios, and of central limit theorems for settings with many small strata, see Abadie and Imbens [2] and Liu, Ren, and Yang [54].

The large-sample approximation just described does not explicitly account for random variation due to the estimation of the propensity score; in particular, the mean term, which is a function of $\widehat{p}_{ki}$ random variables, is instead treated as fixed. To probe the possible role of such variation, we used the M-estimation framework [91] to represent both the propensity score estimates and the test statistic of interest as part of a single multivariate estimation problem and to construct a new variance estimate for the difference between the test statistic and its estimated mean that attempts to incorporate variation in the estimated propensity scores. However, the new variance estimates were frequently near-identical or slightly smaller than those ignoring the propensity score estimation (presumably due to positive correlation between the estimated mean and the test statistic across propensity score fits); in our simulations the associated tests performed almost identically. Given the simpler form and intuitive permutation analogue for the test ignoring propensity score estimation, we focus on this test going forward. More details are provided on the M-estimation approach in Section A.4 of the online appendix.

3.3.3 Estimates and confidence intervals for an additive effect

So far we have discussed covariate-adaptive randomization inference primarily through the lens of testing the sharp null hypothesis of zero effect. It is also possible to use covariate-adaptive randomization inference to construct confidence intervals in settings where the primary goal is estimating a specific causal effect. The exact details of this process depend on the estimand; for purposes of exposition, we focus on estimating a constant additive treatment effect. More formally, we assume that there exists τ such that

$$Y_{ki}(1) = Y_{ki}(0) + \tau \qquad \text{for all } i. \tag{3.5}$$

and we seek to estimate τ and obtain a confidence interval for our estimate.

First, note that for $\tau \neq 0$, the sharp null hypothesis of no effect no longer holds and observed outcomes Y_{ki} are no longer invariant to treatment status, but the transformed outcomes $Y_{ki} - Z_{ki}\tau$ will be invariant, and a randomization test can be constructed by using $Y_{ki} - Z_{ki}\tau$ as input to the test statistic rather than Y_{ki}. This adjustment allows us to test $H_0 : \tau = \tau_0$ for any value $\tau_0 \in \mathbb{R}$. The set of τ_0-values for which the corresponding test does not reject H_0 at level α is a $1 - \alpha$ confidence set for τ [62, 48, §3.5].

To obtain an estimate of τ in model (3.5), we may simply select the value of τ_0 with the largest two-sided p-value (or the median of such values if many share the same p-value). This estimator was previously suggested for use in treatment-control studies by Branson and Bind [10], who noted its connection to the Hodges-Lehmann estimator [37]; for a helpful review of Hodges-Lehmann estimation in the context of randomization inference in observational studies, see Rosenbaum [77, §2.7.2]. Coudin and Dufour [19] also show that in a slightly different context, an estimator based on maximizing the p-value of a permutation tests retains many of the attractive theoretical properties of the traditional Hodges-Lehmann estimator, including median unbiasedness and asymptotic normality, under mild regularity assumptions. Numerical simulations show that this estimator generally performs well and consistently improves substantially on the naïve difference-in-means estimator (see Table 7 in Section A.5 of the online supplement).

Because covariate-adaptive randomization inference induces a discrete distribution for the test statistic, some care is required in inverting the tests to obtain confidence intervals, especially in small samples. In particular, when the distribution's discreteness makes it impossible to construct an interval with coverage exactly $1 - \alpha$, it is important to ensure that intervals are made conservative rather than anticonservative. Luo et al. [56] provide a very helpful practical discussion; while they focus on a uniform design without stratification, the key ideas translate almost without modification to covariate-adaptive randomization inference at least when the difference-in-means statistic of Section 3.3.2 is used. Inverting the test requires repeated tests for a variety of candidate τ-values and can be computationally demanding. However, several shortcuts are possible. In matched pair designs, Monte Carlo draws of the difference-in-means statistic from the null distribution under $H_0 : \tau = \tau_0$ are simple algebraic modifications of Monte Carlo draws from the null distribution under $H_0 : \tau = 0$. In particular, letting Z_{ki}^{orig} represent the original value of treatment in the observed data for subject i in pair k (as opposed to the value Z_{ki} in the draw from the null distribution), we have the following:

$$\begin{aligned} T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}} - \mathbf{Z}_{\mathcal{M}}^{orig}\tau_0) &= \frac{1}{K}\sum_{k=1}^{K}(Z_{k1} - Z_{k2})[Y_{k1} - Y_{k2} - \tau_0(Z_{k1}^{orig} - Z_{k2}^{orig})] \\ &= T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) - \tau_0 \cdot \frac{1}{K}\sum_{k=1}^{K}(Z_{k1} - Z_{k2})(Z_{k1}^{orig} - Z_{k2}^{orig}) \\ &= T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) - \tau_0 \cdot \frac{1}{K}\left[\sum_{k=1}^{K}\mathbf{1}\left\{\mathbf{Z}_k = \mathbf{Z}_k^{orig}\right\} - \sum_{k=1}^{K}\mathbf{1}\left\{\mathbf{Z}_k \neq \mathbf{Z}_k^{orig}\right\}\right] \end{aligned}$$

In summary, as long as we keep track of the number of pairs in each null draw that have been switched relative to the original treatment assignment, we need only take Monte Carlo draws once from the null distribution for $H_0 : \tau = 0$ and can simply transform them as necessary to test any $H_0 : \tau = \tau_0$. Another option not limited to paired designs is to use the large-sample normal approximation to the null distribution of the difference-in-means estimator instead of Monte Carlo draws from the permutation distribution. To invert this large-sample test, one need only evaluate a normal tail probability for each τ_0 considered.

3.4 Assessing the impact of propensity score estimation error

A primary concern in determining the empirical value of covariate-adaptive randomization inference is understanding the impact of errors in estimating the propensity score. To trust the method, we need some guarantee that small deviations between $\widehat{\lambda}(x)$ and true $\lambda(x)$ values result in only small deviations in nominal and actual test size and confidence interval coverage. The following results provide partial reassurance in this regard. To lay the groundwork, let $\widehat{\lambda}_N(\cdot)$ be an estimated propensity score fit on an external sample of size N from the infinite population, and let the quantities $\mathbf{Z}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)$ refer to a separate sample organized into K matched pairs containing $\sum_{k=1}^{K} n_k = n$ total units. Let F_λ represent the true conditional distribution of $\mathbf{Z}$ as a function of the true propensity score $\lambda(\cdot)$ and let $F_{\widehat{\lambda}_N}$ represent the distribution of $\mathbf{Z}$ used to conduct covariate-adaptive inference in practice, based on $\widehat{\lambda}_N(\cdot)$. Furthermore, let $p_{\widehat{\lambda}_N,n}$ be the p-value produced by a nominal level-α covariate-adaptive randomization test of the sharp null hypothesis using $F_{\widehat{\lambda}_N}$. The following result, adapted from Berrett et al. [9] (who consider the slightly different case in which permutations are across the entire dataset and not within matched sets), relates these quantities using the total variation distance of F_λ and $F_{\widehat{\lambda}_N}$.

Theorem 4. *If no unobserved confounding is present and the sharp null hypothesis of no effect is true, then*

$$P(p_{\widehat{\lambda}_N,n} \leq \alpha) \;\leq\; \alpha + d_{TV}\left(F_\lambda, F_{\widehat{\lambda}}\right)$$

where $d_{TV}(P, Q)$ gives the total variation distance between two probability distributions P and Q.

In words, this theorem says that the true type I error of a covariate-adaptive randomization test performed with nominal level α is no larger than α plus the total-variation discrepancy of the distributions implied by the true and estimated propensity score. Intuitively, when the estimated propensity score is close to the true propensity score, this means that the nominal type I error rate is close to the true type I error rate. We formalize this intuition for the case in which the true propensity score obeys a logistic regression model.

Theorem 5. *Suppose that $P(Z = 1 \mid X) = \lambda(X) = \frac{1}{1+\exp(-\beta^T X)}$ and that $\widehat{\lambda}_N$ is obtained by estimating this model using maximum likelihood. Suppose furthermore that N increases with*

the primary sample size n such that $\lim_{n\longrightarrow\infty} n/N = 0$, and suppose the covariates X have compact support. Then under the conditions of Theorem 4,

$$\limsup_{n,N\longrightarrow\infty} P(p_{\widehat{\lambda}_N,n} \leq \alpha) \ \leq \ \alpha.$$

The proof, which uses Pinsker's inequality to bound the total variation distance by a sum of Kullback-Leibler divergences and a Taylor expansion to show that this sum converges to zero, is deferred to Section A.2 of the online appendix.

A natural question is whether similar results can be obtained when the propensity model is not necessarily logistic. Berrett et al. [9] sketch a proof for a result similar to Theorem 5 when the propensity score is estimated nonparametrically using a kernel method; this requires only mild smoothness conditions rather than a correctly-specified model, although the bound on the rate of convergence of the size of the Type I error violation towards zero is much weaker. More generally, one may turn to Theorem 4 directly to explore misspecification: this result provides a bound on Type I error violation whenever the degree of misspecification can be quantified by the total variation distance between estimated and true conditional distributions of treatment. Such metrics for probing robustness appear elsewhere in the literature; for example, Guo et al. [32] assess robustness of inferences to covariate noise by considering a regime in which the conditional distribution of the true covariates given treatment has bounded total variation distance from the conditional distribution of the noise-contaminated covariates.

Another limitation of Theorem 5, which extends also to the nonparametric kernel approach just mentioned, is the asymptotic regime. The issues extend beyond the usual question of whether a particular sample size is large enough to ensure that error is small; here the requirement that the pilot sample used to estimate the propensity score grow at a larger rate than the analysis sample raises similar concerns even for very large samples, since the key question is whether the pilot sample's size sufficiently exceeds that of the analysis sample. Fitting the propensity score in a large independent sample consisting of most of the observed data while reserving a relatively smaller portion for the analysis is uncommon in applied matching studies. However, we note that this approach is natural in settings where treatment and covariate values are readily available but outcomes are expensive or difficult to measure, as when researchers must collect outcomes by administering a test to study subjects [69] or by abstracting medical charts [89]. In Section 3.6, we evaluate the performance of the method by simulation in multiple settings, including some that plausibly adhere to the assumed asymptotic regime and others that likely do not. In general we find only minor difference in performance for tests based on estimated propensity scores fit out-of-sample in samples much larger than the analysis samples versus tests based on propensity scores estimated in-sample on the analysis data, suggesting that the test may be fairly robust to violations of the asymptotic regime given in Theorem 5 in many finite sample settings.

3.5 Covariate-adaptive randomization inference under unobserved confounding

3.5.1 Review of sensitivity analysis framework under exact matching

Results in Sections 3.2 -3.4 have all depended on the absence of unobserved confounding, but in practice it is not plausible that all confounders are observed. To address this issue with the matched randomization inference framework, Rosenbaum [77, §4] presents a method of sensitivity analysis to relax the no unobserved confounding assumption. Specifically, the true probabilities of treatment are restricted as follows:

$$1/\Gamma \leq \frac{\pi(x,u)(1-\pi(x,u'))}{\pi(x,u')(1-\pi(x,u))} \leq \Gamma \quad \text{for all } x,u,u'. \tag{3.6}$$

This is equivalent to the following model for treatment assignment, with arbitrary $\kappa(\cdot)$:

$$\log\left(\frac{\pi(X,U)}{1-\pi(X,U)}\right) = \kappa(X) + \gamma U \quad \text{where } \gamma = \log(\Gamma) \text{ and } 0 \leq U \leq 1. \tag{3.7}$$

To complete the sensitivity analysis, some method is needed to compute worst-case p-values over all possible values of of $\mathbf{U}$ allowed by the model. The method depends on the structure of matched sets formed and the test statistic used; we focus on the approach of Rosenbaum [79] which allows for arbitrary numbers of controls matched to each treated unit and applies to any sum statistic, i.e. any statistic $T(\mathbf{Z},\mathbf{Y}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y})$ for chosen functions f_{ki} of $\mathbf{Y}$, which we denote as the pseudo-responses. As shown in Rosenbaum [79, §5], the difference-in-means statistic, the regression-adjusted test statistics of Rosenbaum [74], and other M-statistics are all members of the family of sum statistics.

Without loss of generality, suppose the n_k units in each matched set k are arranged in increasing order of their pseudo-responses f_{ki} so $f_{k1} \leq f_{k2} \leq \ldots, f_{kn_k}$ for all k, and suppose that we are interested in a one-sided test where larger values of the test statistic will lead to rejection. Let U^+ be the set of all stratified N-tuples $\mathbf{u}$ such that $u_{ki} \in \{0,1\}$ and $u_{k1} \leq u_{k2} \leq \ldots \leq u_{kn_k}$ for all k, and let U^- be the set of all stratified N-tuples $\mathbf{u}$ such that $u_{ki} \in \{0,1\}$ and $u_{k1} \geq u_{k2} \geq \ldots \geq u_{kn_k}$ for all k. Let $\alpha_{unif}(\mathbf{Y},\mathbf{u})$ represent the p-value for the uniform randomization test performed with test statistic $T(\mathbf{Z},\mathbf{Y})$ and the uniform randomization distribution when model (3.7) holds with unobserved confounder $\mathbf{U} = \mathbf{u}$. When matching on the propensity score is exact, Rosenbaum and Krieger [80] showed the following:

$$\min_{\mathbf{u}\in U^-} \alpha_{unif}(\mathbf{Y},\mathbf{u}) \leq \alpha_{unif}(\mathbf{Y},\mathbf{U}) \leq \max_{\mathbf{u}\in U^+} \alpha_{unif}(\mathbf{Y},\mathbf{u}).$$

Although $\mathbf{U}$ is still unknown, this statement allows us to bound its impact on the result of the hypothesis test by searching over a highly structured finite set of candidate $\mathbf{u}$-values. We now show that this approach still works to identify the worst-case p-value when matches are not exact on the propensity score and permutation probabilities are covariate-adaptive.

3.5.2 Sensitivity analysis model under covariate-adaptive randomization inference.

As in the previous section, let $T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})$ be a sum statistic and focus on the case of a one-sided test where large values lead to rejection. We now define $\alpha_{adapt}(\mathbf{Y}, \mathbf{u})$ as the p-value obtained by conducting a covariate-adaptive randomization test (using the true propensity score) when model (3.7) holds with unobserved confounder $\mathbf{U} = \mathbf{u}$.

Theorem 6. *For any* $\mathbf{u} \in [0,1]^n$,

$$\min_{\mathbf{u}' \in U^-} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}') \leq \alpha_{adapt}(\mathbf{Y}, \mathbf{u}) \leq \max_{\mathbf{u}'' \in U^+} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}'').$$

Intuitively, this result says that we can find the maximum (minimum) p-value by placing the maximum (minimum) possible treatment probability on the subjects with the largest pseudo-responses, and the minimum (maximum) treatment probability on those with the smallest pseudo-responses. The full proof, which is partially adapted from results for the exact-matching case in Rosenbaum [77] and from results for weighting estimators in Zhao, Small, and Bhattacharya [101], is deferred to Section A.3 of the online appendix.

While Theorem 6 provides a foundation for sensitivity analysis, two additional issues must be addressed. First, while in the exact-matching case the function $\kappa(X)$ has no bearing on sensitivity analysis (since the $\kappa(X)$ terms cancel within matched pairs), under covariate-adaptive randomization inference some bound or estimate of $\kappa(X)$ is needed to compute sensitivity bounds. In general, the known propensity score $\lambda(X)$ provides information about $\kappa(X)$ as follows:

$$\lambda(X) = E(\pi(X,U) \mid X) = \int_0^1 \pi(X,u) dP(u) = \int_0^1 \frac{1}{1+\exp[-\kappa(X) - \gamma u]} dP(u).$$

When $\gamma = 0$ this gives us a one-to-one mapping between $\kappa(X)$ and $\lambda(X)$ but otherwise such a mapping requires knowledge about the population distribution of U. Fortunately, since $\pi(x,u)$ is increasing in u for all x, for any fixed x we have $\pi(x,0) \leq \lambda(x) \leq \pi(x,1)$. This in turn implies:

$$\kappa(X) \in \left[\log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - \gamma, \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) \right]$$

Note that this bound is tight, since we can make $\kappa(X)$ arbitrarily close to either bound by letting $P(U=0)$ or $P(U=1)$ be arbitrarily close to 1. Therefore we can rewrite model (3.7) in terms of $\lambda(X)$ as:

$$\begin{aligned} \log\left(\frac{\pi(X,U)}{1-\pi(X,U)}\right) &= \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - \gamma V + \gamma U, \qquad U, V \in [0,1]. \\ &= \log\left(\frac{\lambda(X)}{1-\lambda(X)}\right) - \gamma + 2\gamma U', \qquad U' \in [0,1]. \end{aligned} \tag{3.8}$$

The second issue is computational. The bound in Theorem 6 gives us a finite set of possible distributions over which to search to identify the worst-case p-value for a covariate-adaptive randomization test. However, under certain configurations of strata size and number, this set may grow large and complicated so that it is difficult to compute the exact maximum efficiently. Gastwirth, Krieger, and Rosenbaum [29] simplified the problem by showing that asymptotically the overall maximum and minimum p-values for any given Γ are achieved by solving simpler optimization problems separately for each stratum and aggregating the results. Specifically, if $T_k = \sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y}_{\mathcal{M}})$ $\mu_{k\ell}$ is the additive contribution of stratum k to the test statistic and $\mathbf{u}_k$ is the subvector of $\mathbf{u}$ for stratum k, the expectation of T_k must be maximized (minimized) over all binary $\mathbf{u}_k$, and when multiple values $\mathbf{u}_k$ are maximizers (minimizers) the variance of T_k must also be maximized over these. By an argument essentially identical to the one we use to prove Theorem 6, one may show that T_k's expectation is maximized only for $\mathbf{u}_k$ such that $u_{k1} = \ldots = u_{k\ell} = 0$ and $u_{k(\ell+1)} = \ldots, u_{kn_k} = 1$ for some $\ell = 0, 1, \ldots, n_k$, so that it suffices to consider the n_k expectations $\mu_{k\ell}$ and n_k variances $\nu_{k\ell}$ associated with these vectors (for the lower bound similar quantities are used but with a different choice of binary u_{ki} values). Aggregating worst-case expectations and variances across many strata and letting the number of strata go to infinity results in worst-case bounds on the overall p-value. While this result is asymptotic, Rosenbaum [79] later derived a one-step adjustment that renders the bounds conservative in finite samples. Under covariate-adaptive randomization inference and representation (3.8) for the treatment model, the key quantities $\mu_{k\ell}$ and $\nu_{k\ell}$ take on the following values:

$$\mu_{k\ell} = \frac{\sum_{i=1}^{\ell} p_{ki} f_{ki} + \Gamma^2 \sum_{i=\ell+1}^{n_k} p_{ki} f_{ki}}{\sum_{i=1}^{\ell} p_{ki} + \Gamma^2 \sum_{i=\ell+1}^{n_k} p_{ki}} \qquad \nu_{k\ell} = \frac{\sum_{i=1}^{\ell} p_{ki} f_{ki}^2 + \Gamma^2 \sum_{i=\ell+1}^{n_k} p_{ki} f_{ki}^2}{\sum_{i=1}^{\ell} p_{ki} + \Gamma^2 \sum_{i=\ell+1}^{n_k} p_{ki}} - \mu_{k\ell}^2$$

The method for combining these quantities to provide an overall conservative bound on the p-value is identical to that described in Rosenbaum [79].

3.6 Finite-sample performance via simulation

3.6.1 Motivation

The results of Section 3.4 provide some assurance that in extremely large samples, covariate-adaptive randomization inference will closely approximate the true conditional distribution of treatment and outperform uniform randomization inference. However, many matched studies work with relatively small sample sizes, and researchers do not always have strong guarantees that fitted models are well-specified. In what follows we assess the empirical performance of covariate-adaptive randomization inference in small to moderate samples via a simulation, considering cases where models are correctly and incorrectly specified. First, in Section 3.6.2, we consider the performance of covariate-adaptive randomization inference in matched designs compared to more traditional inference procedures for matched studies, with a focus on Type I error control.

A second important question is whether modifying a permutation procedure to account for observed propensity scores obviates the need for matching at all. Indeed, the test proposed by Branson and Bind [10], which considers a similar data setting and relies on the same idea as our test but does not incorporate matching, is also finite-sample valid when conducted with true propensity scores. In Section 3.6.3, we conduct additional simulations to demonstrate that the combination of matching and covariate-adaptive randomization inference offers important benefits in terms of both precision and robustness to misspecification of the propensity model relative to non-uniform permutation in an unmatched study.

3.6.2 Matching with and without covariate-adaptive randomization inference

We create a matrix of covariates X by drawing p vectors of independent standard normal random variables, each vector of length n. p is 2, 5, or 10, and n is either 100 or 1000. The true propensity score is then given by one of the following two functions, one that specifies the logit of treatment as a linear function of the columns of X and another specifying it as a nonlinear function of those columns:

$$\text{logit}\left[P(Z=1 \mid X)\right] = \log\left(\frac{0.3}{0.7}\right) + \Delta \cdot X_1 \tag{3.9}$$

$$\text{logit}\left[P(Z=1 \mid X)\right] = \log\left(\frac{0.3}{0.7}\right) + \frac{\Delta}{\sqrt{265}}\left(X_1 + 4X_1^3\right) \tag{3.10}$$

The functional forms of these models are adapted from those used by Resa and Zubizarreta [70]. The parameter Δ controls the strength of the propensity score signal; we examine two signals, "weak signal" with $\Delta = 0.2$ and "strong signal" with $\Delta = 0.6$. The intercept is chosen to ensure a treated:control ratio in the general neighborhood of 3:7 (since all X_j variables have mean zero over repeated samples), providing a relatively large control pool for pair matching. The scaling factor $1/\sqrt{265}$ is chosen to render the overall signal-to-noise ratio comparable across models for a given value of Δ. For a more detailed look at the distribution of the propensity scores in the treated and control groups under this setup, see Figure 4 in Section A.5 of the online appendix.

Matching is conducted on the joint (X, Z) datasets created by this process. The matching is done using a robust Mahalanobis distance on the columns of X, either with or without a propensity score caliper. When the caliper is used, a propensity score must first be estimated. This is done by fitting a logistic model with linear additive terms using maximum likelihood. Then matches are restricted to occur only between individuals separated by no more than 0.2 sample standard deviations of the fitted propensity scores (computed for the original dataset); if treated units must be excluded from the match to meet this condition the minimal possible number of such exclusions is made.

Finally, outcomes are drawn using a linear model with the same right-hand side as the model used to generate the true propensity score, albeit with no intercept, with $\Delta = 1$,

and with additive independent mean-zero normal errors with variance 4. We conduct randomization inference by reshuffling treatment indicators within pairs uniformly at random, based on propensity scores estimated as described above, or based on true propensity scores. The observed test statistic is compared to the null distribution obtained by 5000 Monte Carlo draws to see whether one-sided tests for a difference greater than zero reject at the 0.05 level. Results are evaluated for both raw outcomes and outcomes adjusted by ordinary least-squares linear regression using the approach of Rosenbaum [74].

In addition, for each combination of simulation parameters we also test the results of inference when the observed test statistic is not generated by the original $\mathbf{Z}$-vector but by a within-match permutation of treatment indicators using distribution (3.2) with the true propensity scores. This is designed to approximate a model in which matched pairs have independent treatment assignments, and eliminates the potential for Z-dependence as discussed in Section 3.2.3.

For datasets of size $n = 100, p = 2$ and $n = 100, p = 5$ we ran each unique combination of the remaining simulation parameters (propensity model, signal strength, caliper indicator, inference method, regression adjustment indicator, and indicator for reshuffled Z-vector in observed test statistic) 8000 times. Type I error rates are calculated as the sample average of rejection indicators over the 8000 draws for each parameter combination. We also ran each combination of parameters 8000 times for datasets of size $n = 1000, p = 10$. We opted not to examine the $n = 100, p = 10$ case because we wanted to focus on cases in which completed matches met standard balance criteria and sample runs suggested that most matches were not successful in balancing all ten covariates, and we studied only $p = 10$ for the much more computationally expensive $n = 1000$ case in order to produce simulation results in a reasonable timeframe.

Figure 3.1 shows the primary results on Type I error rate from the simulations. The most apparent pattern is the high type I error rates for uniform inference on uncalipered matches with the difference-in-means statistic, much higher than the rates for any method that accounts in any additional way for in-pair covariate discrepancies. This is true across every simulation setting shown but especially so at $n = 1000$.

More nuanced are the distinctions among settings using covariate-adaptive inference, calipers, regression-adjusted test statistics, and combinations thereof. In the first three columns of each table, corresponding to settings in which Z-dependence is present, settings with calipers or regression adjusted test statistics tend to perform best, either with or without covariate-adaptive inference; in contrast, settings relying entirely on covariate-adaptive inference tend to do slightly worse or even substantially worse when $n = 1000$. However, when treatment assignments within matched pairs are made independent, eliminating Z-dependence, covariate-adaptive inference alone is generally just as effective as regression adjustment and does better than calipers.

Across the four horizontal groupings in each table, we see different combinations of correctly/incorrectly specified treatment and outcome models. Covariate-adaptive inference with estimated propensity scores tends to fare more poorly when the treatment model is misspecified, even when Z-dependence is not present. Regression adjustment is somewhat

robust to the degree of nonlinearity introduced in this simulation's outcome model. However, when both outcome and treatment models are nonlinear type I error control often fails for $p > 2$. At $n = 1000$ the case with both models misspecified leads to gross violation of type I control under any form of adjustment except oracle covariate-adaptive inference (which uses true, nonlinear treatment probabilities that are unavailable in practice).

All the results shown above focus on the strong propensity signal case ($\Delta = 0.6$), in which matching tends to be substantially more difficult than in the weak signal case. Density plots and type I error rates for the weak-signal case are given in Figures 5 and 6 in Section A.5 of the online appendix. While the overall ordering among approaches remains similar, the size of type I error violations is greatly reduced such that the simple uniform approach often controls Type I error when $p = 2$.

We also conducted two robustness checks to confirm that our results are not sensitive to secondary aspects of our simulation setting. Firstly, we reproduced the Type I error results excluding all simulation runs in which a matched sample failed to achieve absolute standardized differences less than 0.2 on all measured covariates (in case the initial simulations were contaminated by poor matches in some iterations). Secondly, we repeated our simulations with propensity scores fit on independent pilot samples of size 10,000, simulated from the same data-generating process as the analysis sample (a setting more like the one assumed in Theorem 5). These robustness checks all substantiated the patterns observed in the primary simulations (see Figures 7-8 in Section A.5 of the online appendix).

Figure 3.2 describes confidence interval width for simulation settings that achieved approximate Type I control. Whenever a Bonferroni-corrected test against the null hypothesis of a true Type I error rate no more than 0.05 failed to reject, we used the central limit theorem approximation of Section 3.3.2 to create large-sample confidence intervals for an additive effect. For cases with regression-adjusted test statistics, the procedures of Section 3.3.2 were used with residuals from the regression model in place of raw outcomes Y_{ki}. The two main takeaways are the value of regression adjustment for improving precision, a pattern that appears in other parts of the causal inference literature [74, 26, 4], and a cost in precision associated with imposing calipers. This cost comes from reduced sample size, since treated units may be excluded by the caliper leading to fewer matched sets formed. Notably, introducing covariate-adaptive inference never hurt precision among cases with controlled Type I error.

In summary, the simulations emphasize the importance of taking measures of some kind beyond optimal matching itself to control covariate discrepancies within matched pairs and guarantee Type I error control for post-match inference. Moreover, the importance of such measures appears to increase with the size and complexity of the dataset. With respect to Type I error control, the type of correction or adjustment employed seems to matter less than the fact of employing it. Secondly, while covariate-adaptive inference in isolation is slightly less successful than calipers or regression adjustment in many settings, this appears to be largely a product of Z-dependence, since drawing treatment assignments independently across matched pairs tends to erase this gap in most cases. This finding suggests the importance of the specific stochastic model used to justify matched randomization inference,

and the value of further work on understanding Z-dependence and finding effective ways to remove it. Finally, the simulations show that among adjustment strategies that do not rely on fitting some form of outcome model, covariate-adaptive inference holds a small edge over caliper approaches because it does not require a reduction in sample size to fix problems with in-pair covariate discrepancies.

3.6.3 Covariate-adaptive randomization inference with and without matching

We next conduct simulations to articulate the possible benefits of conducting matching followed by covariate-adaptive randomization relative to applying randomization inference directly to unmatched data as in Branson and Bind [10]. First we consider gains to precision in inference. The results of Section 3.3.2 show that the null variance of the difference-in-means statistic under covariate-adaptive randomization inference in a pair-matched study is a function of terms D_k^2 where D_k is the difference between the potential outcomes under control in pair k. While the variance of the difference-in-means statistic is much harder to compute under covariate-adaptive randomization inference under the Bernoulli assignment mechanism of Branson and Bind [10], it tends to depend on the overall dispersion of all outcomes in the study (rather than on the dispersion within matched pairs). As a result, when heterogeneity in the study outcomes is increased but the quality of the pairings remains good, we expect to see bigger gaps in performance between matched and unmatched settings. Accordingly, for this comparison we adopt the "strong signal" version of the simulation setting in Section 3.6.2 but further strengthen the outcome model by multiplying the true signal by a factor of ten before adding noise. We then compare covariate-adaptive randomization inference (with and without z-dependence) to unmatched covariate-adaptive randomization inference as detailed by Branson and Bind [10], assuming a Bernoulli design with independent treatment probabilities equal to the propensity scores. For each design, we examine both the difference-in-means and regression-adjusted test statistic.

Table 3.1 gives Type I errors and confidence interval lengths for the matched and unmatched versions across several parameter combinations (focusing on the case with $n = 100$, $p = 2$, strong propensity signal, and propensity scores estimated in-sample). When well-specified regression models are fit, the matched and unmatched studies show similar performance in the absence of Z-dependence, with slightly more precise confidence intervals coming from the unmatched study; this makes sense because the unmatched study has more data available with which to estimate the outcome model. However, when regression adjustment is absent or based on an incorrect model, the matched design has substantially increased precision relative to the unmatched designs, with confidence intervals less than half as large in some cases. We note also that the Type I errors for the unmatched cases without correctly-specified regression adjustment tend to lie orders of magnitude below 0.05, suggesting that the inference achieved in these cases is punishingly conservative. Evidently, matching offers potential for important precision gains relative to unmatched studies (at the

cost of introducing z-dependence) even when both employ covariate-adaptive inference.

Finally, we probe the possibility that matching confers robustness to incorrect model assumptions using a new simulation setting with a badly misspecified propensity score. The covariate matrix X is generated identically to previous simulations, by drawing p vectors of independent standard normal random variables, but the true propensity score here is given by

$$\text{logit}\left[P(Z = 1 \mid X)\right] = \log\left(\frac{0.2}{0.7}\right) + 1\{X_1 X_2 \geq 0\}, \tag{3.11}$$

Because of the indicator function on the sign of an interaction term $X_1 X_2$, two very distinct values of the propensity score are present, associated with subjects in different quadrants of (X_1, X_2) space, but each covariate is marginally uncorrelated with treatment. The outcomes are drawn from a linear model with a zero intercept, additive noise, and the same function of covariates on the right-hand side. Table 3.2 presents the Type 1 errors for both the matched and unmatched version under the misspecified model (with $n = 100$, $p = 2$, and propensity scores estimated in-sample). The unmatched study exhibits uncontrolled Type I errors due to poor propensity score estimation. However, in the matched setting, Type I errors remain close to 0.05 because matched pairs with small covariate distances tend also to group subjects in the same quadrant of the (X_1, X_2) space, leading to closely-matched propensity scores. Clearly for at least some settings matching confers robustness benefits on covariate-adaptive randomization inference not enjoyed by unmatched studies.

3.6.4 Removing Z-dependence: a small example

The simulations of Sections 3.6.2-3.6.3 showed that Z-dependence poses a problem for settings where the original observations are sampled independently from a population prior to matching. One may argue that matching itself is best understood as a preprocessing or data cleaning step and that the model of independent assignments in matched pairs is as plausible as the model of independently sampling of observations [36], and under such a setting the results presented above without Z-dependence are most relevant for guiding practice. However, a method that can address Z-dependence under independent sampling of observations seems desirable.

With unlimited computing power, it would be trivial to construct such a method. For each draw from the null distribution, one could take the newly-generated vector of treatment assignments $\mathbf{Z}^*$ and repeat the original matching algorithm using it. If the same set of matched pairs is obtained, then $\mathbf{Z}^*$ is a valid draw from the conditional distribution of treatment given $\mathcal{Z}_{\mathcal{M}}$; on the other hand, if a different set of matched pairs is obtained then $\mathbf{Z}^*$ is outside the support of the true conditional distribution.

We apply this procedure in an example generated from the same data-generating process described for the unmatched vs. matched comparisons in Section 3.6.3 but with $n = 30$, $p = 2$, and true propensity scores. In the particular random instance we study, there are 10 treated subjects, and $2^{10} = 1024$ unique permutations of treatment assignment in

the resulting ten matched pairs. Rerunning the matching algorithm with each of these permutations in turn, we find that only 300 lead to the same match. Those rejected looked systematically different than those included, with an average difference-in-means values of -0.40 vs. 1.22 respectively. The original covariate-adaptive test produces a p-value of 0.044 while the reduced-support test gives a p-value of 0.133 which is no longer significant at the 0.05 level. Repeating this analysis a total of 2000 times, we find the estimated Type I error of the original covariate-adaptive procedure to be 0.311 compared to 0.0421 in the reduced-support case. Unfortunately, this strategy does not scale easily to more realistically-sized datasets due to the explosion of the number of possible permutations and the need to compute a match for each. For more discussion, see Section 3.8.

3.7 Performance in case studies

3.7.1 Welders and genetic damage

We now apply the tools developed for covariate-adaptive randomization inference to a real dataset due originally to Costa, Zhitkovich, and Toniolo [18]. This dataset compares the rate of DNA-protein cross-links, a risk factor for gene expression problems, among welders and controls in other occupations. In addition to DNA-protein linkage, age, race, and smoking behavior are measured for all subjects (each of whom is male). Rosenbaum [75] analyzed this data by matching each of the 21 welders to one of the 26 controls. Here we replicate this matching approach and consider the impact of covariate-adaptive randomization inference in place of the uniform randomization inference typically used for pair-matched designs.

Following Rosenbaum, we estimate the propensity score using logistic regression against the three covariates discussed above and match each welder to a control using a robust Mahalanobis distance with a propensity score caliper equal to half the standard deviation of the fitted propensity score values across the dataset. Since it is not possible to match all 21 welders within the caliper, a soft caliper is used in which violations of the caliper are penalized linearly.

As noted by Rosenbaum, pair matching cannot remove all confounding in this data, particularly because only five of the 26 controls are removed by the matching process. Table 3.3 shows the pair differences on each of the three matching variables and the estimated propensity score, with average discrepancies at the bottom. Note that although the average discrepancies are small on all variables (the average difference in age is only 1.5 years), in the large majority of pairs the treated unit has a slightly higher propensity score, several with differences larger than 0.1. In contrast to standard uniform randomization inference, covariate-adaptive randomization inference will pay attention to these differences.

Using the difference in mean DNA-protein linkage between welders and matched controls as a test statistic, we contrast uniform randomization inference structured by the matched pairs with covariate-adaptive inference based on the estimated propensity score. The density plot in Figure 3.3 illustrates the difference between the two null distributions. As the consis-

tent positive sign on the propensity score discrepancies would suggest, the covariate-adaptive randomization distribution is shifted to the right of the uniform randomization distribution, showing how accounting for residual propensity score differences leads us to expect larger values even under the null.

The data was collected based on a hypothesis that welders experience elevated levels of genetic damage compared to controls, so we conduct a one-sided test of the Fisher sharp null hypothesis of no difference in DNA-protein linkage due to welder status. Using the null distributions shown above, this corresponds to calculating the proportion of null draws that exceed the observed value of 0.64. For the uniform distribution this is 0.015, and for the covariate-adaptive distribution it is 0.029. The covariate-adaptive distribution adjusts for the residual propensity score differences in the pairs, which are biased towards treatment for the welders who actually experienced treatment; as such, it recognizes that part of the apparent treatment effect in the uniform test is likely a result of bias and concludes that the weight of evidence in favor of a treatment is weaker. While both tests are significant at the 0.05 level in this case, if researchers had been interested in effects of both signs and had conducted a two-sided test, the covariate-adaptive test does not reject in this case. As such, failure to account properly for propensity score discrepancies in the two-sided case leads to an anticonservative result in which the null hypothesis is rejected when most properly it should not be. When a sensitivity analysis is run, the uniform analysis is similarly more optimistic about evidence for a treatment effect, reporting similar qualitative results for Γ up to 1.09, while the covariate-adaptive analysis allows a maximum value of only 1.05.

3.7.2 Right-heart catheterization and death after surgery

We next conduct covariate-adaptive randomization inference in the right-heart catheterization data of Connors et al. [17]. In this study the effectiveness of the right-heart catheterization (RHC) procedure for improving outcomes of critically ill patients was assessed by matching patients receiving the procedure to those not receiving it and comparing mortality rates. We follow the original authors by fitting a propensity score on 31 variables measured at intake and forming matches only between patients with propensity scores that differ by no more than 0.03. Unlike the authors, who use a greedy matching procedure, we use an optimal matching procedure that minimizes a robust Mahalanobis distance formed from the 31 variables in the propensity score model within propensity score calipers, and also matches exactly on primary disease category. Not all of the 2,184 RHC patients can be matched to distinct controls within the caliper, and in this case we exclude the minimal number of RHC patients necessary for the caliper to be respected. This leaves us with 1,538 matched pairs, which is still a substantial improvement on the match conducted by the authors, which had only 1,008 pairs. Table 3.4 summarizes the effectiveness of the match in removing bias on the observed pre-treatment variables. The numbers shown are standardized treatment-control differences in means for each variable, both before (first column) and after (second column) matching. Only the variables with the 25 largest pre-matching absolute standardized differences are shown (but post-matching standardized differences remain below 0.04 for all

variables not shown). Clearly, although large differences between groups are present in the raw data for numerous variables, notably APACHE risk score (aps1), mean blood pressure (meanbp1), and P/F oxygen ratio (pafi1), matching transforms these into small differences. Of special note is the row showing balance on the estimated propensity score, which shows a reduction from a standardized difference exceeding 1 to a value of only 0.04.

To assess the role of the caliper in influencing the study's results, we construct an alternative matched design using a generalized version of optimal subset matching [81], as implemented in the R package `rcbsubset`. Instead of using a propensity score caliper, this match enforces exact balance on ventiles of the propensity score, ensuring close balance on the propensity score without imposing hard restrictions on the propensity score discrepancy within a matched pair. Treated units are excluded from the match if their inclusion induces the average Mahalanobis distance across matched pairs to cross a specific threshold given by a tuning parameter; we chose a value of the tuning parameter that induces a similar overall sample size as in the calipered match (1,507 matched pairs). Balance on important pre-treatment variables is similar to that in the caliper match, as shown in the third column of Table 3.4. However, there are important differences between the two matches at the level of the pairs, as shown in Table 3.5; the caliper match achieves much greater similarity of propensity scores within pairs, at the cost of achieving slightly reduced similarity on a range of other variables. In practice the caliper match is likely to be more attractive, since it achieves a major improvement in propensity score control at the cost of relatively minor changes in other variables, but both matches are Pareto optimal in the sense of Pimentel and Kelz [66].

The outcome of interest in this study is patient death within thirty days. For each match we conduct four outcome analyses: uniform and covariate-adaptive randomization inference for the difference in means, and uniform and covariate-adaptive randomization inference for a regression-adjusted difference-in-means statistic [74]. Although the outcome is binary, the mortality rate is sufficiently high that an ordinary least-squares fit is reasonable. Table 3.6 summarizes the results of the analysis, providing point estimates, p-values, and estimated confidence intervals from all four approaches, with the point estimates for the covariate-adaptive procedures shifted by the mean of the covariate-adaptive null distribution given in formula (3.3). In addition, results from a sensitivity analysis are reported, giving the largest values of Γ consistent with a rejection of the null hypothesis (the "threshold" Γ).

Results are generally consistent with an increased risk of mortality associated with right-heart catheterization, with all confidence intervals contained in the range 3%-11%. As in the welders data, the covariate-adaptive procedure reports a treatment effect slightly smaller than the uniform procedure; however, in this case the magnitude of the difference is small compared to the overall size of the effect so the methods all agree qualitatively. Notice also that the regression-adjusted test statistics have narrower confidence intervals than the raw risk differences, in line with principles described in Rosenbaum [74]. The regression-adjusted analyses lead to smaller estimates of the treatment effect which explains their greater sensitivity to unmeasured bias. The caliper matching has more stable threshold Γ values across analyses, consistent with the tighter control of the propensity score it achieves.

For the subset match, the covariate-adaptive procedure tends to increase the threshold Γ despite having smaller treatment effects; this is because it also lowers the variance of the estimator and leads to tighter confidence intervals.

3.8 Discussion

Uniform randomization inference for matched studies relies, often at least in part implicitly, on a model in which unobserved confounders are absent, in which propensity scores are matched exactly, and (depending on the sampling model) in which the matched sets selected are conditionally independent of the original treatment vector. Substantial failures of these assumptions lead to substantial problems with Type I error control and require some form of correction. Covariate-adaptive randomization provides such correction by altering permutation probabilities in the randomization test based on estimated propensity scores. Relative to the naïve analysis for the difference-in-means statistic, covariate-adaptive randomization tends to restore approximate control of Type I error in many settings and constitutes an attractive option alongside approaches based on regression-adjusted test statistics and matching calipers. Furthermore, the combination of matching and covariate-adaptive randomization inference offers value not provided by incorporating estimated propensity scores into permutation procedures for unmatched studies. Specifically, matched designs can enjoy greatly improved precision relative to unmatched studies with similar non-uniform permutation procedures, and are also more robust to misspecification of the propensity score, although they are subject to an additional source of Type I error violations (z-dependence) to which unmatched studies are not. We note also that the sensitivity analysis guarantees we develop in Section 3.5 for matched designs do not appear to extend easily to unmatched studies, due to the greater complexity of the conditional distribution of the treatment variable.

The applied examples show that covariate-adaptive randomization inference need not change the qualitative results of an observational study, especially when the study designer has given careful attention to propensity score discrepancies. In these settings the covariate-adaptive procedure can still be a productive robustness check to help build confidence that lingering propensity score discrepancies are not corrupting the study's key findings.

Covariate-adaptive inference raises several interesting future directions for theory and methods development. First, the theoretical guarantee given in this work is limited both by a strong restriction on the relative sample sizes of the analysis sample and of a hypothetical pilot sample used to fit the propensity score. Clarifying whether this assumption can be relaxed and whether same-sample estimation of some kind, such as cross-fitting, could be applied instead would be a valuable contribution. For the difference-in-means estimator, one likely challenge is that the bias of the covariate-adaptive randomization distribution based on the estimated propensity score may not shrink to zero at a strictly faster rate than the variance when propensity-fitting and analysis samples are of a similar order, so that preserving valid inference may require widening confidence intervals or reducing the

significance threshold α to account for the bias.

Secondly, covariate-adaptive randomization inference offers opportunities to develop stronger model-based guidance about the relative importance of the propensity score and the prognostic score in the construction of matched designs. Currently many design objectives proposed as bases for constructing matched sets, including mean balance, caliper matching, and multivariate distance matching have a heuristic element in the sense that it is not clear for which class of overall models for treatment and outcome they provide optimal results. While Kallus [41] lays out a helpful overarching framework under sampling-based inference, unifying descriptions are not present for randomization-based inferences in matching. Covariate-adaptive inference provides important progress towards this goal by clarifying the impact of inexact propensity score matching on the operating characteristics of the ultimate estimate and associated test. Power analysis and design sensitivity calculations [76] based on the covariate-adaptive model, if developed, would provide valuable design-stage insight about how to properly use the propensity score in constructing the matches, and could provide more definitive guidance to users selecting among many Pareto optimal matches [66].

Thirdly, the evidence provided by our simulation study suggests that Z-dependence may play a limited but non-trivial role in failures of type I error control for matched randomization tests. While from a technical standpoint it may be sufficient to assume the matched pairs are generated independently by some model (as opposed to assuming a model on the original observations), this makes it difficult to develop formal understanding of how choices about matching influence inference. In particular, some statistical model on the original subjects of the study is needed to obtain the strong model-based guidance about the proper construction of matched sets discussed in the previous paragraph. Z-dependence differs fundamentally from propensity score discrepancies in the sense that it alters the support of the randomization distribution, not just the values of nonzero permutation probabilities, and new methods. The procedure outlined in Section 3.6.4, in which each permutation is checked to see if it produces the same match, works well for very small study sizes. Improving its computational performance seems possible; in particular, it should not be necessary to create an optimal match for each separate permutation, merely to check whether the current match is optimal or not, and this task may be easier to accomplish efficiently. We view developing such checks for commonly-used matching procedures as a promising future research direction.

Finally, while our development has focused exclusively on sharp null hypotheses under which both potential outcomes are known for all individuals, there is substantial practical interest in testing weak null hypotheses which allow for unknown heterogenous treatment effects and merely restrict the averages of such effects. Several threads of recent work have demonstrated that under mild conditions inference strategies developed to test sharp null hypotheses may also be valid tests for appropriately-chosen weak null hypotheses. An important task is to determine whether these ideas work for covariate-adaptive randomization inference. Notably, Caughey et al. [11] showed that permutation tests of a sharp null may be reinterpreted as tests of a weak null describing maxima of treatment effects rather than averages. Since Caughey et al. place almost no restrictions on the permutation distribution and

only mild conditions on the test statistic (all satisfied by the difference-in-means estimator we consider above), we anticipate that this approach should apply almost without modification to covariate-adaptive randomization tests. Fogarty (2020, 2022) describes sharp-null-style permutation tests and sensitivity analyses that are valid for the more traditional weak null hypothesis of zero average effect. These tests generally require studentized test statistics, and are much harder to construct for designs with matched sets containing more than two individuals. As such, more substantial extensions of our existing work are needed to see if the same ideas work under covariate-adaptive randomization inference.

Statement on conflicts of interest and data availability

The authors report no conflicts of interest. The welders data due originally to Costa, Zhitkovich, and Toniolo [18] may be obtained via the R package `DOS2`, which is freely available at the Comprehensive R Archive Network (CRAN). The right heart catheterization data due originally to Connors et al. [17] may be obtained via the R package `RBestMatch`. Although `RBestMatch` is not currently hosted by CRAN, archived versions of the package that contain the data are still freely available through the CRAN website at `https://cran.r-project.org/src/contrib/Archive/RBestMatch/`.

Specification			Type I Error			Confidence interval length		
Z-model	Y-model	Regression?	Unmatched	Matched. Z-dependence?		Unmatched	Matched. Z-dependence?	
				With	Without		With	Without
Linear	Linear	No	0.000	0.284	0.045	8.68	3.41	3.43
Linear	Linear	Yes	0.053	0.166	0.047	1.81	5.73	2.01
Linear	Nonlinear	No	0.001	0.037	0.051	7.70	2.11	3.22
Linear	Nonlinear	Yes	0.041	0.014	0.052	4.76	3.57	2.03
Nonlinear	Linear	No	0.000	0.258	0.056	8.85	3.39	5.22
Nonlinear	Linear	Yes	0.048	0.265	0.055	1.78	5.89	1.90
Nonlinear	Nonlinear	No	0.008	0.044	0.064	7.96	2.09	5.16
Nonlinear	Nonlinear	Yes	0.217	0.040	0.056	4.85	3.50	1.89

Table 3.1: Simulation results comparing matched and unmatched studies employing covariate-adaptive randomization inference. The first three columns describe distinct simulation settings (for $n = 100$ and $p = 2$ with other parameters given as described in Section 3.6.3). The middle three columns show Type I error computed as the proportion of rejections in a nominal level-0.05 test against a larger alternative across 5000 iterations; the first column gives the value for covariate-adaptive randomization inference in the full dataset without matching following Branson and Bind [10], and the second two give values for covariate-adaptive randomization after matching (with and without Z-dependence) as detailed in Section 3.6.2. The final three columns show the average of two-sided confidence interval length (computed by inverting the two-sided version of the corresponding test) computed over the same set of iterations.

n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.055	0.057	0.056	0.054	0.055	0.055	Yes	Yes
		0.041	0.045	0.044	0.054	0.059	0.058	No	Yes
		0.051	0.058	0.058	0.051	0.060	0.060	Yes	No
		0.064	0.083	0.110	0.054	0.073	0.093	No	No
Linear	Nonlinear	0.048	0.048	0.048	0.052	0.051	0.051	Yes	Yes
		0.041	0.040	0.038	0.055	0.051	0.050	No	Yes
		0.044	0.048	0.049	0.052	0.056	0.056	Yes	No
		0.055	0.061	0.078	0.053	0.058	0.066	No	No
Nonlinear	Linear	0.051	0.050	0.051	0.051	0.051	0.050	Yes	Yes
		0.028	0.029	0.028	0.048	0.048	0.046	No	Yes
		0.051	0.052	0.052	0.051	0.053	0.052	Yes	No
		0.060	0.056	0.088	0.048	0.048	0.065	No	No
Linear	Linear	0.050	0.049	0.050	0.053	0.054	0.052	Yes	Yes
		0.042	0.041	0.041	0.052	0.054	0.051	No	Yes
		0.049	0.050	0.052	0.054	0.056	0.057	Yes	No
		0.068	0.065	0.089	0.054	0.053	0.065	No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.062	0.071	0.070	0.046	0.052	0.052	Yes	Yes
		0.042	0.053	0.048	0.054	0.072	0.063	No	Yes
		0.044	0.067	0.069	0.048	0.078	0.078	Yes	No
		0.057	0.074	0.131	0.052	0.082	0.115	No	No
Linear	Nonlinear	0.057	0.055	0.056	0.053	0.052	0.051	Yes	Yes
		0.046	0.044	0.040	0.047	0.050	0.045	No	Yes
		0.042	0.054	0.055	0.052	0.073	0.073	Yes	No
		0.062	0.060	0.100	0.048	0.056	0.077	No	No
Nonlinear	Linear	0.058	0.055	0.056	0.050	0.049	0.048	Yes	Yes
		0.031	0.033	0.029	0.048	0.052	0.044	No	Yes
		0.042	0.049	0.050	0.050	0.064	0.063	Yes	No
		0.060	0.052	0.112	0.054	0.058	0.091	No	No
Linear	Linear	0.058	0.056	0.054	0.050	0.048	0.048	Yes	Yes
		0.042	0.045	0.039	0.046	0.051	0.044	No	Yes
		0.041	0.046	0.050	0.048	0.070	0.072	Yes	No
		0.073	0.062	0.127	0.046	0.047	0.078	No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.052	0.080	0.080	0.048	0.074	0.074	Yes	Yes
		0.030	0.073	0.068	0.048	0.118	0.109	No	Yes
		0.041	0.098	0.100	0.052	0.116	0.118	Yes	No
		0.072	0.234	0.521	0.050	0.184	0.449	No	No
Linear	Nonlinear	0.048	0.049	0.050	0.051	0.051	0.051	Yes	Yes
		0.046	0.045	0.044	0.055	0.055	0.052	No	Yes
		0.037	0.059	0.063	0.051	0.081	0.084	Yes	No
		0.097	0.138	0.334	0.054	0.083	0.224	No	No
Nonlinear	Linear	0.044	0.044	0.043	0.051	0.050	0.050	Yes	Yes
		0.025	0.025	0.022	0.049	0.050	0.044	No	Yes
		0.042	0.054	0.059	0.051	0.067	0.070	Yes	No
		0.109	0.103	0.397	0.048	0.046	0.254	No	No
Linear	Linear	0.050	0.050	0.050	0.052	0.051	0.051	Yes	Yes
		0.046	0.047	0.043	0.049	0.050	0.046	No	Yes
		0.048	0.061	0.065	0.050	0.067	0.070	Yes	No
		0.140	0.133	0.436	0.050	0.050	0.225	No	No

Figure 3.1: Type I error results for uniform and covariate-adaptive inference across multiple simulation settings. Each of the three tables corresponds to a separate dataset size. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models. Numbers give type I error rates with colors associated to their magnitude; triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05 (under a Bonferroni correction scaled to the number of results across the entire figure), with a large upper triangle indicating a positive z-statistic and a small lower triangle indicating a negative z-statistic.

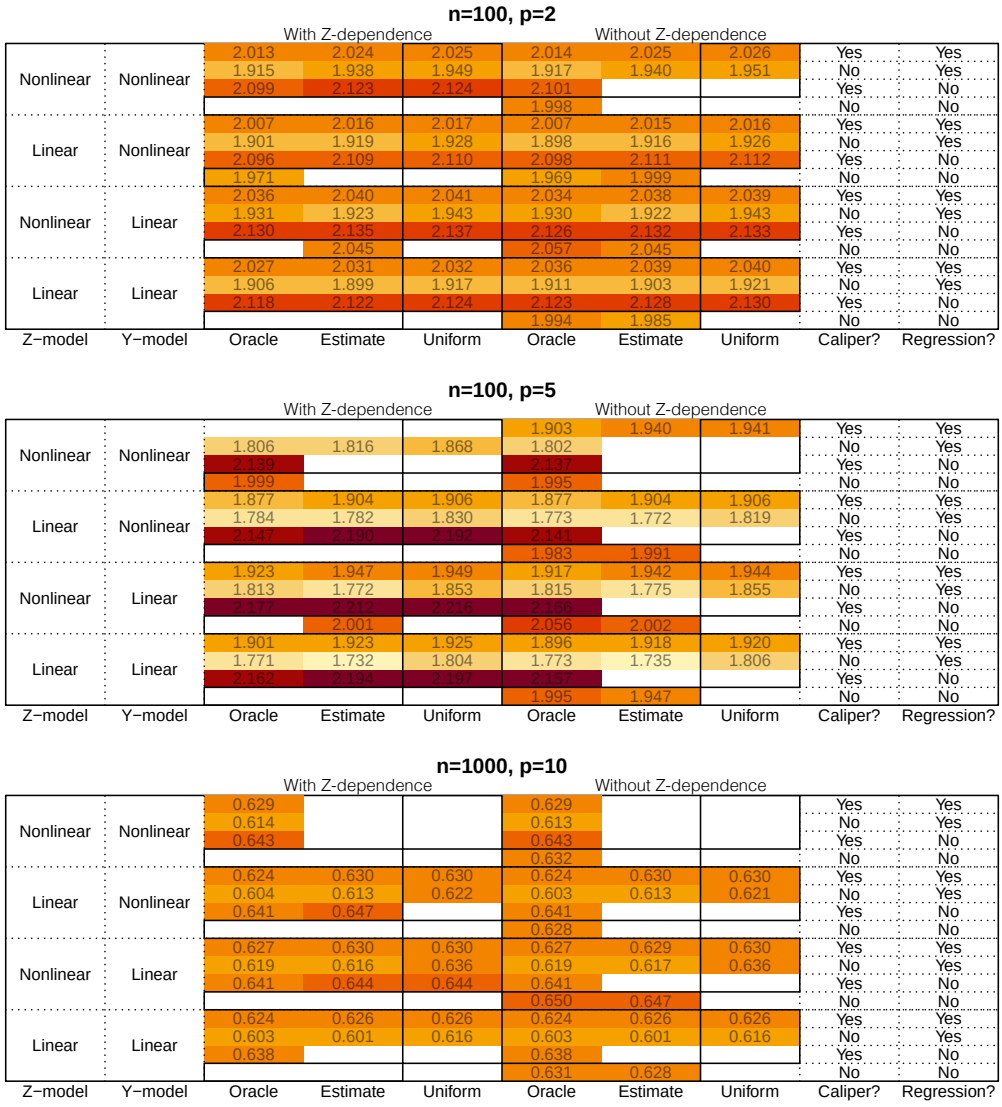

Figure 3.2: Average confidence interval length for uniform and covariate-adaptive inference across multiple simulation settings for cases with approximate control of type I error at 0.05 or less. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models.

	Type I Error		
Regression?	Unmatched	Matched, with Z-dependence	Matched, no Z-dependence
No	0.135	0.072	0.064
Yes	0.139	0.059	0.058

Table 3.2: Simulation results comparing the robustness of matched and unmatched studies employing covariate-adaptive randomization inference to misspecification of the propensity score. Data are generated from a process with misspecification in both the propensity score and the outcome model as described in Section 3.6.4. Type I errors are computed as the proportion of rejections in a nominal level-0.05 test against a larger alternative across 5000 iterations.

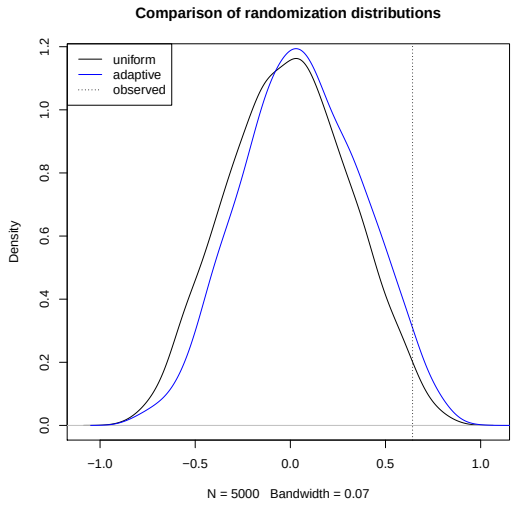

Figure 3.3: Smoothed densities for the uniform randomization distribution of the difference in means statistic and the covariate-adaptive randomization distribution in the welders dataset, with the value of the observed statistic.

	African-American	Age	Smoker	PS
Pair 1	0.00	-6.00	0.00	0.12
Pair 2	0.00	-3.00	0.00	0.05
Pair 3	0.00	4.00	0.00	0.05
Pair 4	0.00	-8.00	0.00	0.16
Pair 5	0.00	0.00	0.00	0.00
Pair 6	0.00	3.00	0.00	-0.06
Pair 7	0.00	-3.00	0.00	0.06
Pair 8	0.00	-2.00	0.00	0.17
Pair 9	0.00	3.00	0.00	-0.06
Pair 10	0.00	-5.00	0.00	0.06
Pair 11	0.00	-3.00	0.00	0.06
Pair 12	0.00	-5.00	0.00	0.10
Pair 13	0.00	-5.00	0.00	0.11
Pair 14	0.00	-5.00	0.00	0.10
Pair 15	0.00	0.00	0.00	0.00
Pair 16	0.00	-4.00	0.00	0.08
Pair 17	0.00	1.00	0.00	-0.02
Pair 18	0.00	1.00	0.00	-0.01
Pair 19	-1.00	7.00	0.00	0.04
Pair 20	0.00	-5.00	0.00	0.10
Pair 21	0.00	4.00	0.00	0.05
Average	-0.05	-1.48	0.00	0.06

Table 3.3: Treated- control differences in matched pairs in the welders dataset.

	Before Matching	Caliper Match	Subset Match
Est. Propensity Score	1.254	0.002	0.033
Acute Physiology Score	0.501	0.016	0.068
Mean Blood Pressure	-0.455	-0.026	-0.058
PaO2/(.01*FiO2)	-0.433	0.001	0.030
Serum Creatinine	0.270	0.021	0.044
Hematocrit	-0.269	0.001	-0.008
Weight (kg)	0.256	0.002	-0.017
PaCO2	-0.249	0.004	-0.013
MOSF w/Sepsis (secondary)	0.230	-0.036	-0.014
Albumin	-0.230	-0.018	-0.021
Predicted Survival	-0.198	-0.031	-0.049
MOSF w/Sepsis (primary)	0.172	0.000	0.000
Respiration Rate	-0.165	-0.002	-0.011
Heart Rate	0.147	0.027	0.000
Bilirubin	0.145	-0.013	0.010
Heart disease	0.139	0.011	0.014
MOSF w/Malignancy (secondary)	-0.135	0.009	-0.017
Respiratory Disease	-0.128	-0.007	-0.020
Serum pH	-0.120	-0.002	-0.044
Missing ADL Score	0.117	-0.001	0.005
SUPPORT Coma Score	-0.110	0.045	0.022
Neurological disease	-0.108	0.006	0.001
Serum Sodium	-0.092	-0.011	-0.001
Years of Education	0.091	0.022	0.024
Sepsis	0.091	0.012	0.017

Table 3.4: Standardized differences in means before matching and under two different matches for the 25 variables with largest initial imbalance in the right-heart catheterization dataset.

	Caliper Match	Subset Match
Avg. Propensity Score Discrepancy	0.016	0.166
Max Propensity Score Discrepancy	0.030	0.672
Prop. Matched Exactly, Male Sex	0.351	0.297
Avg. Discrepancy, Age	13.469	10.892
Avg. Discrepancy, Mean Blood Pressure	27.809	25.506
Avg. Discrepancy, Heart Rate	32.677	28.858
Avg. Discrepancy, Respiration Rate	11.614	10.426
Avg. Discrepancy, PaO2/(.01*FiO2)	76.651	75.882

Table 3.5: Matched pair quality under two different matches for selected variables the right-heart catheterization dataset.

	Caliper match		Subset match	
	Uniform	Covariate-Adaptive	Uniform	Covariate-Adaptive
Risk difference	0.074	0.073	0.068	0.062
Lower conf. limit	0.043	0.041	0.038	0.034
Upper conf. limit	0.106	0.104	0.098	0.089
Threshold Γ	1.12	1.12	1.11	1.12
Risk difference, OLS-adjusted	0.051	0.051	0.055	0.054
Lower conf. limit, OLS-adjusted	0.022	0.022	0.027	0.028
Upper conf. limit, OLS-adjusted	0.081	0.081	0.083	0.079
Threshold Γ, OLS-adjusted	1.06	1.06	1.08	1.09

Table 3.6: Outcome analysis for 30-day mortality in the right-heart catheterization data. Results are shown for both the caliper match and the subset match, with and without regression adjustment, and using both uniform and covariate-adaptive inference.

List of figure captions

1. Type I error results for uniform and covariate-adaptive inference across multiple simulation settings. Each of the three tables corresponds to a separate dataset size. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models. Numbers give type I error rates with colors associated to their magnitude; triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05 (under a Bonferroni correction scaled to the number of results across the entire figure), with a large upper triangle indicating a positive z-statistic and a small lower triangle indicating a negative z-statistic.

2. Average confidence interval length for uniform and covariate-adaptive inference across multiple simulation settings for cases with approximate control of type I error at 0.05 or less. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models.

3. Smoothed densities for the uniform randomization distribution of the difference in means statistic and the covariate-adaptive randomization distribution in the welders dataset, with the value of the observed statistic.

Bibliography

[1] Alberto Abadie and Guido W Imbens. "Bias-corrected matching estimators for average treatment effects". In: *Journal of Business & Economic Statistics* 29.1 (2011), pp. 1–11.

[2] Alberto Abadie and Guido W Imbens. "Large sample properties of matching estimators for average treatment effects". In: *Econometrica* 74.1 (2006), pp. 235–267.

[3] Abhineet Agarwal et al. "Pcs-uq: Uncertainty quantification via the predictability-computability-stability framework". In: *arXiv preprint arXiv:2505.08784* (2025).

[4] Joseph Antonelli et al. "Doubly robust matching estimators for high dimensional confounding adjustment". In: *Biometrics* 74.4 (2018), pp. 1171–1179.

[5] Susan Athey, Guido W Imbens, and Stefan Wager. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.4 (2018), pp. 597–623.

[6] Peter C Austin. "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples". In: *Statistics in medicine* 28.25 (2009), pp. 3083–3107.

[7] Peter C Austin and Elizabeth A Stuart. "Optimal full matching for survival outcomes: a method that merits more widespread use". In: *Statistics in medicine* 34.30 (2015), pp. 3949–3967.

[8] Mike Baiocchi. "Methodologies for observational studies of health care policy". PhD thesis. University of Pennsylvania, 2011.

[9] Thomas B Berrett et al. "The conditional permutation test for independence while controlling for confounders". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1 (2020), pp. 175–197.

[10] Zach Branson and Marie-Abèle Bind. "Randomization-based inference for Bernoulli trial experiments and implications for observational studies". In: *Statistical methods in medical research* 28.5 (2019), pp. 1378–1398.

[11] Devin Caughey et al. "Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects". In: *arXiv preprint arXiv:2101.09195* (2021).

[12] Ambarish Chattopadhyay and José R Zubizarreta. "On the implied weights of linear regression for causal inference". In: *Biometrika* 110.3 (2023), pp. 615–629.

[13] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[14] Victor Chernozhukov et al. "Long story short: Omitted variable bias in causal machine learning". In: *The Review of Economics and Statistics* (2021), pp. 1–45.

[15] Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. "sensemakr: Sensitivity analysis tools for OLS in R and Stata". In: *Observational Studies* 10.2 (2024), pp. 93–127.

[16] Carlos Cinelli and Chad Hazlett. "Making sense of sensitivity: Extending omitted variable bias". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.1 (2020), pp. 39–67.

[17] Alfred F Connors et al. "The effectiveness of right heart catheterization in the initial care of critically III patients". In: *Jama* 276.11 (1996), pp. 889–897.

[18] Max Costa, Anatoly Zhitkovich, and Paolo Toniolo. "DNA-protein cross-links in welders: molecular implications". In: *Cancer research* 53.3 (1993), pp. 460–463.

[19] Élise Coudin and Jean-Marie Dufour. "Finite-sample generalized confidence distributions and sign-based robust estimators in median regressions with heterogeneous dependent errors". In: *Econometric Reviews* 39.8 (2020), pp. 763–791.

[20] Alicia Curth and Mihaela Van Der Schaar. "In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation". In: *International conference on machine learning*. PMLR. 2023, pp. 6623–6642.

[21] Rajeev H Dehejia and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". In: *Journal of the American Statistical Association* 94.448 (1999), pp. 1053–1062.

[22] Peng Ding and Tirthankar Dasgupta. "A potential tale of two-by-two tables from completely randomized experiments". In: *Journal of the American Statistical Association* 111.513 (2016), pp. 157–168.

[23] Jacob Dorn, Kevin Guo, and Nathan Kallus. "Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding". In: *Journal of the American Statistical Association* just-accepted (2024), pp. 1–23.

[24] Raaz Dwivedi et al. "Stable discovery of interpretable subgroups via calibration in causal studies". In: *International Statistical Review* 88 (2020), S135–S178.

[25] Ronald A Fisher. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.

[26] Colin B Fogarty. "Regression-assisted inference for the average treatment effect in paired experiments". In: *Biometrika* 105.4 (2018), pp. 994–1000.

[27] Colin B Fogarty. "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies". In: *Journal of the American Statistical Association* 115.531 (2020), pp. 1518–1530.

[28] Colin B Fogarty. "Testing weak nulls in matched observational studies". In: *Biometrics* (2022).

[29] Joseph L Gastwirth, Abba M Krieger, and Paul R Rosenbaum. "Asymptotic separability in sensitivity analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.3 (2000), pp. 545–555.

[30] Alan S Gerber, Donald P Green, and Christopher W Larimer. "Social pressure and voter turnout: Evidence from a large-scale field experiment". In: *American Political Science Review* 102.1 (2008), pp. 33–48.

[31] Kevin Guo and Dominik Rothenhäusler. "On the statistical role of inexact matching in observational studies". In: *Biometrika* 110.3 (2023), pp. 631–644.

[32] Wenshuo Guo et al. "Partial Identification with Noisy Covariates: A Robust Optimization Approach". In: *arXiv preprint arXiv:2202.10665* (2022).

[33] Ben B Hansen. "Full matching in an observational study of coaching for the SAT". In: *Journal of the American Statistical Association* 99.467 (2004), pp. 609–618.

[34] Ben B Hansen. *Propensity score matching to recover latent experiments: diagnostics and asymptotics*. Tech. rep. 486. University of Michigan, 2009.

[35] Chad Hazlett. "Angry or weary? How violence impacts attitudes toward peace among Darfurian refugees". In: *Journal of Conflict Resolution* 64.5 (2020), pp. 844–870.

[36] Daniel E Ho et al. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". In: *Political analysis* 15.3 (2007), pp. 199–236.

[37] Joseph L Hodges Jr and Erich L Lehmann. "Estimates of location based on rank tests". In: *The Annals of Mathematical Statistics* 34.2 (1963), pp. 598–611.

[38] Paul W Holland. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[39] Melody Huang and Samuel D Pimentel. "Variance-based sensitivity analysis for weighting estimators result in more informative bounds". In: *arXiv preprint arXiv:2208.01691* (2022).

[40] Siddharth Jain et al. "Risk of Parkinson's disease after anaesthesia and surgery". In: *British Journal of Anaesthesia* 128.4 (2022), e268–e270.

[41] Nathan Kallus. "Generalized optimal matching methods for causal inference." In: *J. Mach. Learn. Res.* 21 (2020), pp. 62–1.

[42] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. "Removing hidden confounding by experimental grounding". In: *Advances in neural information processing systems* 31 (2018).

[43] Joseph DY Kang and Joseph L Schafer. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical Science* 22.4 (2007), pp. 523–529.

[44] Alan B Krueger. "Experimental estimates of education production functions". In: *The Quarterly Journal of Economics* 114.2 (1999), pp. 497–532.

[45] Sören R. Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10 (Mar. 2019), pp. 4156–4165. DOI: 10.1073/pnas.1804597116. URL: https://www.pnas.org/doi/full/10.1073/pnas.1804597116 (visited on 09/04/2024).

[46] Soonwoo Kwon and Liyang Sun. "Estimating treatment effects under bounded heterogeneity". In: *arXiv preprint arXiv:2510.05454* (2025).

[47] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American Economic Review* (1986), pp. 604–620.

[48] Erich Leo Lehmann. *Testing statistical hypotheses*. John Wiley & Sons, 1959.

[49] Lihua Lei and Emmanuel J Candès. "Conformal inference of counterfactuals and individual treatment effects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5 (2021), pp. 911–938.

[50] Yan Leng and Drew Dimmery. "Calibration of heterogeneous treatment effects in randomized experiments". In: *Information Systems Research* 35.4 (2024), pp. 1721–1742.

[51] Liang Li and Tom Greene. "A weighting analogue to pair matching in propensity score analysis". In: *The international journal of biostatistics* 9.2 (2013), pp. 215–234.

[52] Xinran Li and Peng Ding. "General forms of finite population central limit theorems with applications to causal inference". In: *Journal of the American Statistical Association* 112.520 (2017), pp. 1759–1769.

[53] Kung-Yee Liang and Scott L Zeger. "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1 (1986), pp. 13–22.

[54] Hanzhong Liu, Jiyang Ren, and Yuehan Yang. "Randomization-based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2 K Factorial Experiments". In: *Journal of the American Statistical Association* (2022), pp. 1–15.

[55] Hanzhong Liu and Yuehan Yang. "Regression-adjusted average treatment effect estimates in stratified randomized experiments". In: *Biometrika* 107.4 (2020), pp. 935–948.

[56] Xiaokang Luo et al. "Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.4 (2021), pp. 777–797.

[57] Kathryn R Mahaffey, Robert P Clickner, and Catherine C Bodurow. "Blood organic mercury and dietary mercury intake: National Health and Nutrition Examination Survey, 1999 and 2000". In: *Environmental health perspectives* 112.5 (2004), p. 562.

[58] Divyat Mahajan et al. "Empirical analysis of model selection for heterogeneous causal effect estimation". In: *International Conference on Learning Representations*. Vol. 2024. 2024, pp. 26673–26702.

[59] Rahul Mukerjee, Tirthankar Dasgupta, and Donald B Rubin. "Using standard tools from finite population sampling to improve causal inference for complex experiments". In: *Journal of the American Statistical Association* 113.522 (2018), pp. 868–881.

[60] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. "Realcause: Realistic causal inference benchmarking". In: *arXiv preprint arXiv:2011.15007* (2020).

[61] Whitney K Newey. "The asymptotic variance of semiparametric estimators". In: *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382.

[62] Jerzy Neyman. "Outline of a theory of statistical estimation based on the classical theory of probability". In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767 (1937), pp. 333–380.

[63] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319.

[64] Nicole E Pashley, Guillaume W Basse, and Luke W Miratrix. "Conditional as-if analyses in randomized experiments". In: *Journal of Causal Inference* 9.1 (2021), pp. 264–284.

[65] Samuel D Pimentel and Yaxuan Huang. "Covariate-adaptive randomization inference in matched designs". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86.5 (2024), pp. 1312–1338.

[66] Samuel D Pimentel and Rachel R Kelz. "Optimal tradeoffs in matched designs comparing US-trained and internationally trained surgeons". In: *Journal of the American Statistical Association* 115.532 (2020), pp. 1675–1688.

[67] Samuel D Pimentel et al. "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons". In: *Journal of the American Statistical Association* 110.510 (2015), pp. 515–527.

[68] Samuel D Pimentel et al. "Optimal matching approaches in health policy evaluations under rolling enrolment". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183.4 (2020), pp. 1411–1435.

[69] June Machover Reinisch et al. "In utero exposure to phenobarbital and intelligence deficits in adult men". In: *Jama* 274.19 (1995), pp. 1518–1525.

[70] Maria de los Angeles Resa and José R Zubizarreta. "Evaluation of Subset Matching Methods and Forms of Covariate Balance". In: *Statistics in Medicine* 35.27 (2016), pp. 4961–4979.

[71] Zachary T Rewolinski and Bin Yu. "PCS Workflow for Veridical Data Science in the Age of AI". In: *arXiv preprint arXiv:2508.00835* (2025).

[72] James M Robins and Naisyin Wang. "Inference for imputation estimators". In: *Biometrika* 87.1 (2000), pp. 113–124.

[73] Paul R Rosenbaum. "Conditional permutation tests and the propensity score in observational studies". In: *Journal of the American Statistical Association* 79.387 (1984), pp. 565–574.

[74] Paul R Rosenbaum. "Covariance adjustment in randomized experiments and observational studies". In: *Statistical Science* 17.3 (2002), pp. 286–327.

[75] Paul R Rosenbaum. *Design of Observational Studies*. New York, NY: Springer, 2010.

[76] Paul R Rosenbaum. "Design sensitivity and efficiency in observational studies". In: *Journal of the American Statistical Association* 105.490 (2010), pp. 692–702.

[77] Paul R Rosenbaum. *Observational Studies*. New York, NY: Springer, 2002.

[78] Paul R Rosenbaum. "Optimal matching for observational studies". In: *Journal of the American Statistical Association* 84.408 (1989), pp. 1024–1032.

[79] Paul R Rosenbaum. "Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels". In: *The Annals of Applied Statistics* 12.4 (2018), pp. 2312–2334.

[80] Paul R Rosenbaum and Abba M Krieger. "Sensitivity of two-sample permutation inferences in observational studies". In: *Journal of the American Statistical Association* 85.410 (1990), pp. 493–498.

[81] Paul R. Rosenbaum. "Optimal Matching of an Optimally Chosen Subset in Observational Studies". In: *Journal of Computational and Graphical Statistics* 21.1 (2012), 57–71.

[82] Donald B Rubin. "Comment on 'Randomization analysis of experimental data: The Fisher randomization test'". In: *Journal of the American Statistical Association* 75.371 (1980), pp. 591–593.

[83] Donald B Rubin. "Using propensity scores to help design observational studies: application to the tobacco litigation". In: *Health Services and Outcomes Research Methodology* 2.3 (2001), pp. 169–188.

[84] Fredrik Sävje. "On the inconsistency of matching without replacement". In: *Biometrika* (2021).

[85] Alejandro Schuler et al. "A comparison of methods for model selection when estimating individual treatment effects". In: *arXiv preprint arXiv:1804.05146* (2018).

[86] Azeem M Shaikh and Panos Toulis. "Randomization tests in observational studies with staggered adoption of treatment". In: *Journal of the American Statistical Association* 116.536 (2021), pp. 1835–1848.

[87] Seonho Shin. "Evaluating the effect of the matching grant program for refugees: An Observational Study Using Matching, Weighting, and the Mantel-Haenszel Test". In: *Journal of Labor Research* 43.1 (2022), pp. 103–133.

[88] Jeffrey H Silber et al. "Alzheimer's dementia after exposure to anesthesia and surgery in the elderly: a matched natural experiment using appendicitis". In: *Annals of surgery* (2020).

[89] Jeffrey H Silber et al. "Multivariate matching and bias reduction in the surgical outcomes study". In: *Medical care* (2001), pp. 1048–1064.

[90] Dan Soriano et al. "Interpretable sensitivity analysis for balancing weights". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 186.4 (2023), pp. 707–721.

[91] Leonard A Stefanski and Dennis D Boos. "The calculus of M-estimation". In: *The American Statistician* 56.1 (2002), pp. 29–38.

[92] Elizabeth A Stuart. "Matching methods for causal inference: A review and a look forward". In: *Statistical Science* 25.1 (2010), pp. 1–21.

[93] Zhiqiang Tan. "A distributional approach for causal inference using propensity scores". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1619–1637.

[94] Tiffany M Tang et al. "A simplified MyProstateScore2. 0 for high-grade prostate cancer". In: *Cancer Biomarkers* 42.1 (2025), p. 18758592241308755.

[95] Getayeneh Antehunegn Tesema et al. "Estimating the impact of birth interval on under-five mortality in east african countries: a propensity score matching analysis". In: *Archives of Public Health* 81.1 (2023), p. 63.

[96] Rebecca L Thornton. "The demand for, and impact of, learning HIV status". In: *American Economic Review* 98.5 (2008), pp. 1829–1863.

[97] Ryan J Tibshirani et al. "Conformal prediction under covariate shift". In: *Advances in neural information processing systems* 32 (2019).

[98] Bin Yu and Rebecca L Barter. *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press, 2024.

[99] Bin Yu and Karl Kumbier. "Veridical data science". In: *Proceedings of the National Academy of Sciences* 117.8 (2020), pp. 3920–3929.

[100] Yao Zhang and Qingyuan Zhao. "What is a randomization test?" In: *Journal of the American Statistical Association* just-accepted (2023), pp. 1–29.

[101] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2019).

[102] Qingyuan Zhao, Dylan S Small, and Paul R Rosenbaum. "Cross-screening in observational studies that test many hypotheses". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1070–1084.

[103] José R Zubizarreta. "Using mixed integer programming for matching in an observational study of kidney failure after surgery". In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1360–1371.

Appendix A

Supplementary Material for Chapter 1

A.1 Simulation Setups and Data Examples

A.1.1 Synthetic Data

We use the four simulation setups (A–D) from Nie and Wager [63], which are designed to vary the complexity of the CATE function, the baseline outcome surface, and (in the original paper) the propensity score.

Data-generating process. In all setups, the observed outcome is generated as

$$Y_i = b(X_i) + \left(Z_i - \tfrac{1}{2}\right)\tau(X_i) + \sigma\,\varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \tag{A.1}$$

where $b(\cdot)$ is the baseline (prognostic) function, $\tau(\cdot)$ is the CATE, and $\sigma = 0.5$ throughout. The number of covariates is $p = 6$. Treatment assignment follows $Z_i \overset{iid}{\sim}$ Bernoulli(0.5), i.e., a balanced RCT. The true individual-level CATE under this parameterization is $\tau(X_i)$ by construction.

Modification from the original paper. Nie and Wager [63] designed setups B, C, and D with non-uniform propensity scores to study observational-data performance. We replace these with $e \equiv 0.5$ (balanced treatment) throughout, reflecting the RCT focus of the present paper. The baseline functions $b(\cdot)$ and CATE functions $\tau(\cdot)$ are kept exactly as specified in the original.

The four setups.

- **Setup A.** Covariates $X_i \overset{iid}{\sim}$ Uniform$[0,1]^6$.

$$b(x) = \sin(\pi x_1 x_2) + 2(x_3 - \tfrac{1}{2})^2 + x_4 + \tfrac{1}{2}x_5,$$
$$\tau(x) = \tfrac{1}{2}(x_1 + x_2).$$

The CATE is a smooth, bounded function of two covariates. The baseline surface is moderately complex.

- **Setup B.** Covariates $X_i \overset{iid}{\sim} \mathcal{N}(0, I_6)$.

$$b(x) = \max(0,\, x_1 + x_2,\, x_3) + \max(0,\, x_4 + x_5),$$
$$\tau(x) = x_1 + \log\bigl(1 + e^{x_2}\bigr).$$

The CATE is unbounded and varies nonlinearly; the baseline is piecewise-linear. Designed to favour flexible learners such as the T-learner or X-learner.

- **Setup C.** Covariates $X_i \overset{iid}{\sim} \mathcal{N}(0, I_6)$.

$$b(x) = 2\log\bigl(1 + e^{x_1+x_2+x_3}\bigr),$$
$$\tau(x) = 1.$$

The treatment effect is *constant* (no heterogeneity); a correctly specified constant model is oracle-optimal. The complex baseline surface tests whether estimators are misled by prognostic variation.

- **Setup D.** Covariates $X_i \overset{iid}{\sim} \mathcal{N}(0, I_6)$.

$$b(x) = \tfrac{1}{2}[\max(x_1 + x_2 + x_3,\, 0) + \max(x_4 + x_5,\, 0)]\,,$$
$$\tau(x) = \max(x_1 + x_2 + x_3,\, 0) - \max(x_4 + x_5,\, 0).$$

The CATE can be positive, negative, or zero depending on which linear combination dominates; the function shares structure with the baseline, posing a signal-separation challenge.

Simulation parameters. We vary the training sample size across $n_{\text{train}} \in \{500, 1{,}000, 2{,}000, 5{,}000\}$, with 100 independent replications per configuration. Each replication also generates a separate held-out test set: $n_{\text{test}} = 500$ for the two smaller sizes and $n_{\text{test}} = 1{,}000$ for the two larger sizes. Screening uses a 75%/25% block-stratified train/validation split. We retain $K = 3$ methods after screening and run $B = 500$ bootstrap replicates for uncertainty quantification. The number of calibration subgroups is $L = \max(10,\, \lfloor n_{\text{train}}/100 \rfloor)$, targeting approximately 100 observations per subgroup (giving $L \in \{10, 10, 20, 50\}$ for the four sample sizes), at nominal level $\alpha = 0.05$. The candidate library consists of 14 estimators: the 12 meta-learner variants (S/T/X/R-learner × lasso/ridge/forest) plus `causal_forest` and `bart` (each in a computationally lighter configuration suitable for bootstrap reuse). Because the true $\tau(X_i)$ is available, oracle MSE is computed on the test set for evaluation; it is not used for screening.

A.1.2 Real-world RCT Data

We evaluate the PCS procedure on four benchmark RCT datasets. Because the true individual CATE is unknown in real data, evaluation relies on the oracle-free metrics (Cal, WCal, R-risk). Dataset sizes and variable definitions are as follows.

- **LaLonde / NSW Job Training.** The National Supported Work (NSW) demonstration randomly assigned economically disadvantaged workers to a job-training programme [47]. We use the experimental subsample from Dehejia and Wahba [21], excluding the PSID/CPS observational comparison groups.
  - Treatment Z: received job training (binary).
  - Outcome Y: 1978 earnings (USD).
  - Covariates X ($p = 8$): age, years of education, Black, Hispanic, married, no high-school degree, 1974 earnings, 1975 earnings.
  - Sample: $n = 445$ (185 treated, 260 control); empirical propensity $\hat{e} \approx 0.416$.
  - Source: R package `Matching` (`data(lalonde)`).
- **Tennessee STAR / Kindergarten Class Size.** Project STAR randomly assigned students and teachers to small ($\leq$ 17) or regular-size (22–25) classes within schools [44]. We use kindergarten students with complete math and reading scores, restricting to the small-vs-regular-class comparison.
  - Treatment Z: assigned to small class (binary).
  - Outcome Y: sum of kindergarten math and reading scaled scores.
  - Covariates X ($p = 3$): female, African-American, free lunch eligible.
  - Sample: $n = 3{,}733$ (1,733 small, 2,000 regular); empirical propensity $\hat{e} \approx 0.464$.
  - Source: R package `AER` (`data(STAR)`); also used in Kallus, Puli, and Shalit [42].
- **Get-out-the-Vote (GOTV) / Social Pressure Experiment.** Gerber, Green, and Larimer [30] randomly mailed registered voters in Michigan before the 2006 primary election. We compare the "Neighbors" arm (showing the household's own past voting record alongside neighbours') against the pure control group.
  - Treatment Z: received the social-pressure mailing (binary).
  - Outcome Y: voted in the 2006 primary (binary).
  - Covariates X ($p = 4$): female, age (2006 minus birth year), voted in 2004 primary, household size.
  - Sample: $n = 229{,}444$ (38,201 treated, 191,243 control); empirical propensity $\hat{e} \approx 0.166$.

  - Source: `social.csv` from Imai's `qss` GitHub repository.

- **Thornton HIV Incentives / Malawi Field RCT.** Thornton [96] randomly offered small financial incentives to individuals in Malawi who had previously been tested for HIV, encouraging them to collect their test results.
  - Treatment Z: received any financial incentive (binary).
  - Outcome Y: collected HIV test result at a voluntary counselling and testing centre (binary).
  - Covariates X ($p = 4$): distance to VCT centre, age, HIV-positive status in 2004, village ID.
  - Sample: $n = 2{,}825$ after removing incomplete records (2,211 treated, 614 control); empirical propensity $\hat{e} \approx 0.780$.
  - Source: R package `causaldata` (`data(thornton_hiv)`).

A.1.3 Semi-synthetic Data

The RealCause framework. Neal, Huang, and Raghupathi [60] introduced the *RealCause* benchmarking library for causal inference. The key idea is to fit a generative model to a real observational dataset and then draw new samples from it. Each sample consists of covariate vectors X, along with *both* potential outcomes $(Y(0), Y(1))$, so the true individual CATE $\tau(X) = Y(1) - Y(0)$ is known by construction. This combines the realistic covariate distribution and outcome complexity of a real dataset with the benefit of a known ground truth, which purely observational benchmarks cannot provide.

- For each dataset, Neal, Huang, and Raghupathi [60] provide 100 pre-computed CSV samples (indices 0–99), each drawn from the fitted generative model.
- The generative models are fitted to match the marginal distributions and dependence structure of the underlying real data.
- The resulting samples are therefore *not* RCTs: treatment assignment in the pre-computed CSVs reflects the observational propensity.

Our adaptation: re-randomized RCT. To evaluate PCS in the RCT setting, we re-randomize treatment assignment in every sample:

- For each pre-computed sample, we draw $n_{\text{train}} + n_{\text{test}}$ rows without replacement, then independently assign $Z_i \overset{iid}{\sim} \text{Bernoulli}(0.5)$ to each split.
- The observed outcome is constructed as $Y_i = Z_i \cdot Y_i(1) + (1 - Z_i) \cdot Y_i(0)$, using the potential outcomes from the generative model.

- This yields a balanced RCT with known CATE, while preserving the realistic covariate structure and treatment-effect heterogeneity of the underlying data.
- We use $n_{\text{train}} = 1{,}000$ and $n_{\text{test}} = 500$, with $B = 500$ bootstrap replicates.

Datasets. We use the three RealCause DGPs currently available in the library:

- **LaLonde/CPS** (`lalonde_cps`). Based on the LaLonde NSW job-training experiment merged with the Current Population Survey (CPS) comparison group. Covariates include age, education, race, and pre-intervention earnings. Outcome is 1978 earnings. The treatment-effect distribution in the original observational data is highly skewed.
- **LaLonde/PSID** (`lalonde_psid`). The same NSW experiment merged with the Panel Study of Income Dynamics (PSID) comparison group, yielding a larger and more heterogeneous control pool. Covariate and outcome definitions are identical to the CPS variant. Known to have substantial selection bias in the observational assignment mechanism.
- **Twins** (`twins`). Based on a twins birth dataset, which records outcomes for twin pairs born in the United States between 1989 and 1991. Treatment is low birth weight (an indicator for the lighter twin), outcome is one-year infant mortality. Covariates capture birth characteristics and maternal factors. The CATE is heterogeneous and varies with birth risk profile.

Appendix B

Supplementary Material for Chapter 2

B.1 Individual weight error

Theorem 7 (Error in implied regression weights from omitting U)**.** *Let w_i^* denote the weights from regression with the unobserved covariate U_i. The error in the implied weights from omitting U_i can be expressed as:*

$$w_i^* - w_i = \frac{1}{m_{Z_i}} \left((U_i - \bar{U}_{Z_i}) - (X_i - \bar{X}_{Z_i})^\top v_{Z_i} \right) \Delta_{Z_i} \tag{B.1}$$

where m and v are functions of the covariance between X and U:

- *for LR (ATE),*

$$v_1 = v_0 = v = \hat{\beta}_{U \sim X|Z}, \quad \Delta_1 = -\Delta_0 = (\bar{U}_c - \bar{U}_t) - (\bar{X}_c - \bar{X}_t)^\top v$$
$$m_1 = m_0 = \sum_{i=1}^{n} \left[(U_i - \bar{U}_{Z_i}) - (X_i - \bar{X}_{Z_i})^\top v \right]^2$$

- *for ILR (ATE),*

$$v_z = \hat{\beta}_{U \sim X|Z=z}, \quad \Delta_1 = (\bar{U} - \bar{U}_t) - (\bar{X} - \bar{X}_t)^\top v_1, \ \Delta_0 = (\bar{U} - \bar{U}_c) - (\bar{X} - \bar{X}_c)^\top v_0$$
$$m_z = \sum_{i:Z_i=z} \left[(U_i - \bar{U}_z) - (X_i - \bar{X}_z)^\top v_z \right]^2$$

- *for ILR (ATT),*

$$v_0 = \hat{\beta}_{U \sim X|Z=0}, \quad \Delta_1 = 0, \ \Delta_0 = (\bar{U}_t - \bar{U}_c) - (\bar{X}_t - \bar{X}_c)^\top v_0$$
$$m_0 = \sum_{i:Z_i=0} \left[(U_i - \bar{U}_0) - (X_i - \bar{X}_0)^\top v_0 \right]^2$$

Notes:

1. This expression reveals that the error is driven by the residual imbalance in U that cannot be explained by the observed covariates $\mathbf{X}$. However, the process of residualizing U slightly differs between LR and ILR.

2. In addition, the imbalance that is considered is the mean imbalance, which means if $(\bar{U}_c - \bar{U}_t), (\bar{\mathbf{X}}_c - \bar{\mathbf{X}}_t)^\top \mathbf{v} = 0$ (i.e., no residual mean imbalance), even if other moments of U are imbalanced across treatment and control, $w^* = w$.

3. Potential diagnosis: visualizing how individual weights could be affected by a potential confounder U, or visualizing how an observed covariate affects the weights (like which covariate leads to large/negative weights)

B.2 Population Perspective on L_2 Sensitivity Model

In the main text, all sensitivity analyses are conducted conditional on the realized sample, with perturbations defined over the implied weight vector w. In this appendix, we provide a population-level formulation of the framework using Riesz representers ([61], [13]). This perspective clarifies the underlying estimand and shows that the finite-sample results arise as plug-in analogues of population quantities.

Let P denote the joint distribution of (X, Z, U, Y), and suppose observations are drawn i.i.d. from P. For regression estimators, the population analogue of the implied weights is given by a Riesz representer. In the simple linear regression (LR) setting, by the Frisch–Waugh–Lovell theorem, the population coefficient on Z in the linear projection of Y on $(1, X, Z)$ can be written as

$$\tau_{\mathrm{LR}}(P) = \mathbb{E}_P[\alpha_P(X, Z)Y],$$

where

$$\alpha_P(X, Z) = \frac{Z - \Pi_P[Z \mid 1, X]}{\mathbb{E}_P[(Z - \Pi_P[Z \mid 1, X])^2]},$$

and $\Pi_P[Z \mid 1, X]$ denotes the $L_2(P)$ projection of Z onto the linear span of $(1, X)$. If the regression is augmented with an unobserved confounder U, the corresponding oracle representer becomes

$$\tau^*_{\mathrm{LR}}(P) = \mathbb{E}_P[\alpha^*_P(X, Z, U)Y], \qquad \alpha^*_P(X, Z, U) = \frac{Z - \Pi_P[Z \mid 1, X, U]}{\mathbb{E}_P[(Z - \Pi_P[Z \mid 1, X, U])^2]}.$$

Under this formulation, the natural population analogue of the L_2 perturbation is the $L_2(P)$ distance between representers:

$$\|\alpha^*_P - \alpha_P\|^2_{L_2(P)} = \mathbb{E}_P\left[(\alpha^*_P(X, Z, U) - \alpha_P(X, Z))^2\right].$$

Thus, the sensitivity restriction can be interpreted as constraining the population-level discrepancy between the oracle and observed representers. This perspective is closely related

to the variance-based sensitivity model of [39], which bounds the distributional difference between the weights computed with and without an omitted confounder. The key difference is that their framework is formulated for weighting estimators, where the weights are inverse-propensity or balancing weights, whereas our framework leaves the representer unspecified. It can therefore be applied to the implied weights generated by other procedures, including simple and interacted linear regression.

In the restricted model, this discrepancy admits a natural parameterization via a population analogue of R_w^2:

$$R^2_{w,\text{pop}} := 1 - \frac{\mathbb{E}_P[\alpha_P(X,Z)^2]}{\mathbb{E}_P[\alpha^*_P(X,Z,U)^2]},$$

which measures the fraction of variation in the oracle representer attributable to the omitted confounder. The corresponding population bias bound is

$$|\tau^*(P) - \tau(P)| \leq \|Y^\perp\|_{L_2(P)} \cdot \sqrt{\mathbb{E}_P[\alpha_P(X,Z)^2]} \cdot \sqrt{\frac{R^2_{w,\text{pop}}}{1 - R^2_{w,\text{pop}}}},$$

where $Y^\perp = Y - \Pi_P(Y \mid 1, X, Z)$ denotes the residual from projecting Y onto the model space.

This formulation highlights that our finite-sample sensitivity analysis corresponds to a plug-in approximation of a population problem defined in the space $L_2(P)$, where both the estimand and the sensitivity parameter are expressed in terms of the geometry of the Riesz representer.

B.3 Generalization to more general regression estimators

B.3.1 Weighted Least Squares (WLS)

Many estimators beyond OLS also admit an implied-weight representation. In particular, the weighted least squares analogs studied by Chattopadhyay and Zubizarreta [12], including simple weighted linear regression and interacted weighted linear regression, and the bias-corrected doubly robust estimator (WLS with IPW as the base weights), can all be written in the form

$$\hat{\tau} = w^\top Y,$$

where w is the vector of signed implied weights determined by the observed covariates and the chosen base weights. Moreover, these implied weights solve a quadratic program that balances covariate means toward an estimator-specific target profile while remaining close to the base weights. Thus the L_2 sensitivity framework developed in the main text extends directly once the corresponding implied weight vector is identified. Accordingly, the unrestricted L_2 sensitivity model for WLS is

$$\sigma(c) := \{w^* : \|w^* - w\|_2^2 \leq c\}.$$

The resulting bias bound is immediate.

Proposition 4 (Unrestricted L_2 sensitivity bound for WLS)**.** *Let $\hat{\tau} = w^\top Y$ be any estimator that admits an implied-weight representation, including simple weighted linear regression and interacted weighted linear regression, and the bias-corrected doubly robust estimator. If the oracle target weights satisfy $w^* \in \sigma(c)$, then*

$$|\hat{\tau}^* - \hat{\tau}| = |(w^* - w)^\top Y| \leq \sqrt{c}\,\|Y\|_2.$$

If some coordinates of the weight vector are fixed by construction, the same argument applies after restricting both w and w^* to the coordinates that may vary.

In practice, c can be informally benchmarked by omitting observed covariates from the WLS procedure and recomputing the implied weights, yielding perturbation magnitudes of the form $\|w^{(-j)} - w\|_2^2$. This parallels the unrestricted benchmarking discussion in Section 3 for OLS.

The restricted model for WLS, in the omitted-variable-bias style, is more subtle because the base weights may themselves depend on the observed covariates and therefore may also change when an unobserved confounder U is omitted. Without imposing further structure on how those base weights are generated, it is therefore unclear how to define the restricted oracle class.

B.4 Proofs

B.4.1 Proof of Theorem 1

Proof. By definition,

$$\hat{\tau}^* - \hat{\tau} = (w^* - w)^\top Y = (\Delta w)^\top Y.$$

Under the restricted sensitivity model

$$\{w^* :\ ||w^* - w||_2^2 \leq f(\Gamma),\ \text{and } \exists U \text{ s.t. } w(X, U) = w^*\}$$

both w and w^* are implied weight vectors. Hence they satisfy the same balancing equations:

$$\mathbf{1}^\top w = \mathbf{1}^\top w^* = 0,\ X^\top w = X^\top w^* = 0,\ Z^\top w = Z^\top w^* = 1.$$

Subtracting gives

$$\mathbf{1}^\top \Delta w = 0,\ X^\top \Delta w = 0,\ Z^\top \Delta w = 0.$$

Therefore Δw is orthogonal to the column space of $[\,\mathbf{1}, X, Z\,]$, so by the Cauchy-Schwarz inequality, we have

$$\hat{\tau}^* - \hat{\tau} = (\Delta w)^\top Y = (\Delta w)^\top M_{[\mathbf{1},X,Z]} Y = (\Delta w)^\top Y^{\perp\{\mathbf{1},X,Z\}} \leq \|Y^{\perp\{\mathbf{1},X,Z\}}\|_2 \|\Delta w\|_2.$$

By the definition of R_w^2, we have

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2}.$$

Therefore, substituting the above equation into the bias expression, we get

$$|\hat{\tau}^* - \hat{\tau}| \leq \|Y^{\perp\{\mathbf{1},X,Z\}}\|_2 \, \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}}.$$

For sharpness, let $r := Y^{\perp\{\mathbf{1},X,Z\}}$. If $r = 0$, then the bound is zero and there is nothing to prove. Suppose $r \neq 0$, and define

$$U := r + aZ$$

for some scalar $a \in \mathbb{R}$. Since $r \perp X$ and $r \perp Z$, we have $r \perp Z^{\perp X}$ and

$$U^{\perp X} = r + aZ^{\perp X}.$$

Therefore, by FWL,

$$w^* = \frac{Z^{\perp\{X,U\}}}{\|Z^{\perp\{X,U\}}\|_2^2} = \frac{M_{U^{\perp X}} Z^{\perp X}}{\|M_{U^{\perp X}} Z^{\perp X}\|_2^2},$$

where

$$M_{U^{\perp X}} Z^{\perp X} = Z^{\perp X} - \frac{(U^{\perp X})^\top Z^{\perp X}}{\|U^{\perp X}\|_2^2} U^{\perp X} = \frac{\|r\|_2^2}{\|r\|_2^2 + a^2 \|Z^{\perp X}\|_2^2} Z^{\perp X} - \frac{a\|Z^{\perp X}\|_2^2}{\|r\|_2^2 + a^2 \|Z^{\perp X}\|_2^2} r,$$

using $U^{\perp X} = r + aZ^{\perp X}$ and $r \perp Z^{\perp X}$. Therefore,

$$w^* = \frac{Z^{\perp X}}{\|Z^{\perp X}\|_2^2} - \frac{a}{\|r\|_2^2} r,$$

which gives

$$w^* = w - \frac{a}{\|r\|_2^2} r,$$

so

$$\Delta w = w^* - w = -\frac{a}{\|r\|_2^2} r, \qquad \|\Delta w\|_2^2 = \frac{a^2}{\|r\|_2^2}.$$

Now choose

$$a = \pm \|r\|_2 \, \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}},$$

with sign chosen so that Δw is positively collinear with r. Then

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2},$$

so the sensitivity constraint is attained with equality. Since $\Delta w \propto r$, equality also holds in Cauchy-Schwarz, and hence

$$|\hat{\tau}^* - \hat{\tau}| = \|Y^{\perp\{\mathbf{1},X,Z\}}\|_2 \, \|w\|_2 \sqrt{\frac{R_w^2}{1 - R_w^2}}.$$

Thus the bound is sharp.

For the unrestricted model, if $\|\Delta w\|_2^2 \leq c$, then a direct application of Cauchy–Schwarz gives

$$|\hat{\tau}^* - \hat{\tau}| = |(\Delta w)^\top Y| \leq \|\Delta w\|_2 \, \|Y\|_2 \leq \sqrt{c} \, \|Y\|_2.$$

This proves the unrestricted bound.

□

B.4.2 Proof of Theorems 2 and 3

Proof. We treat ILR-ATE and ILR-ATT in turn.

ILR (ATE). Let $\Delta w := w^* - w$. Under the restricted sensitivity model, both w and w^* are implied-weight vectors from ILR for the ATE. Therefore, within each treatment group $z \in \{0, 1\}$, they satisfy the same balancing equations:

$$\sum_{i:Z_i=z} w_i = \sum_{i:Z_i=z} w_i^* = 1, \qquad \sum_{i:Z_i=z} w_i X_i = \sum_{i:Z_i=z} w_i^* X_i = \bar{X}.$$

Subtracting gives

$$\sum_{i:Z_i=z} \Delta w_i = 0, \qquad \sum_{i:Z_i=z} \Delta w_i X_i = 0, \qquad z \in \{0, 1\}.$$

Hence Δw_t is orthogonal to the column space of $[\mathbf{1}, X_t]$ and Δw_c is orthogonal to the column space of $[\mathbf{1}, X_c]$.

Using the signed implied-weight representation,

$$\hat{\tau}^* - \hat{\tau} = (w^* - w)^\top Y = \Delta w_t^\top Y_t + \Delta w_c^\top Y_c.$$

Projecting within each group gives

$$\Delta w_t^\top Y_t = \Delta w_t^\top Y_t^{\perp\{\mathbf{1},X_t\}}, \qquad \Delta w_c^\top Y_c = \Delta w_c^\top Y_c^{\perp\{\mathbf{1},X_c\}}.$$

Therefore

$$\begin{aligned}
|\hat{\tau}^* - \hat{\tau}| &\leq |\Delta w_t^\top Y_t^{\perp\{\mathbf{1},X_t\}}| + |\Delta w_c^\top Y_c^{\perp\{\mathbf{1},X_c\}}| \\
&\leq \sqrt{\|\Delta w_t\|_2^2 + \|\Delta w_c\|_2^2} \sqrt{\|Y_t^{\perp\{\mathbf{1},X_t\}}\|_2^2 + \|Y_c^{\perp\{\mathbf{1},X_c\}}\|_2^2} \\
&= \|\Delta w\|_2 \sqrt{\|Y_t^{\perp\{\mathbf{1},X_t\}}\|_2^2 + \|Y_c^{\perp\{\mathbf{1},X_c\}}\|_2^2}.
\end{aligned}$$

By the definition of R_w^2 under the restricted model,

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2}.$$

Substituting yields the restricted-model bound in Theorem 2. Under the unrestricted model,

$$|\hat{\tau}^* - \hat{\tau}| = |\Delta w^\top Y| \leq \|\Delta w\|_2 \, \|Y\|_2 \leq \sqrt{c} \, \|Y\|_2,$$

For sharpness of the restricted ATE bound, let

$$r_t := Y_t^{\perp\{\mathbf{1}, X_t\}}, \qquad r_c := Y_c^{\perp\{\mathbf{1}, X_c\}}.$$

If both residual vectors are zero, the bound is zero and there is nothing to prove. Otherwise, by Theorem 7, the perturbation within each treatment group is proportional to the residualized omitted confounder within that group. Hence we may choose an omitted confounder U whose residualized values within the treated and control groups are proportional to r_t and r_c, respectively, with the proportionality constants chosen so that

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2}.$$

With this choice, Δw_t and Δw_c are positively collinear with r_t and r_c, respectively, so equality holds in the groupwise Cauchy–Schwarz step above. Therefore the restricted ATE bound is sharp within the oracle class.

ILR (ATT). Under ILR-ATT, treated implied weights are fixed so

$$\Delta w_i = 0, \qquad Z_i = 1.$$

For controls, both w and w^* satisfy

$$\sum_{i:Z_i=0} w_i = \sum_{i:Z_i=0} w_i^* = 1, \qquad \sum_{i:Z_i=0} w_i X_i = \sum_{i:Z_i=0} w_i^* X_i = \bar{X}_t,$$

so

$$\sum_{i:Z_i=0} \Delta w_i = 0, \qquad \sum_{i:Z_i=0} \Delta w_i X_i = 0.$$

Hence Δw_c is orthogonal to the column space of $[\mathbf{1}, X_c]$ and

$$\hat{\tau}^* - \hat{\tau} = \Delta w_c^\top Y_c = \Delta w_c^\top Y_c^{\perp\{\mathbf{1}, X_c\}}.$$

Applying Cauchy-Schwarz,

$$|\hat{\tau}^* - \hat{\tau}| \leq \|\Delta w_c\|_2 \, \|Y_c^{\perp\{\mathbf{1}, X_c\}}\|_2 \leq \|\Delta w\|_2 \, \|Y_c^{\perp\{\mathbf{1}, X_c\}}\|_2.$$

Substituting

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2}$$

yields the restricted-model bound in Theorem 3. Under the unrestricted model,

$$|\hat{\tau}^* - \hat{\tau}| = |\Delta w_c^\top Y_c| \leq \|\Delta w_c\|_2 \, \|Y_c\|_2 \leq \sqrt{c} \, \|Y_c\|_2.$$

For sharpness of the restricted ATT bound, let $r_c := Y_c^{\perp\{\mathbf{1}, X_c\}}$. If $r_c = 0$, the bound is zero and there is nothing to prove. Otherwise, Theorem 7 shows that the control-group perturbation Δw_c is proportional to the residualized omitted confounder in the control group, while treated weights are fixed. We may therefore choose U so that its residualized values among controls are proportional to r_c, with the scale chosen to satisfy

$$\|\Delta w\|_2^2 = \|w\|_2^2 \frac{R_w^2}{1 - R_w^2}.$$

Then Δw_c is positively collinear with r_c, so equality holds in Cauchy–Schwarz and the restricted ATT bound is sharp within the oracle class.

□

B.4.3 Proof of Proposition 1

Proof. For LR, by the Frisch–Waugh–Lovell theorem, the coefficient on Z can be written as

$$\hat{\tau} = \frac{(Z^{\perp X})^\top Y}{\|Z^{\perp X}\|_2^2} = w^\top Y, \qquad w = \frac{Z^{\perp X}}{\|Z^{\perp X}\|_2^2}.$$

Likewise, if the regression is augmented with an omitted confounder U, the oracle coefficient is

$$\hat{\tau}^* = \frac{(Z^{\perp\{X,U\}})^\top Y}{\|Z^{\perp\{X,U\}}\|_2^2} = (w^*)^\top Y, \qquad w^* = \frac{Z^{\perp\{X,U\}}}{\|Z^{\perp\{X,U\}}\|_2^2}.$$

Therefore,

$$\|w\|_2^2 = \frac{1}{\|Z^{\perp X}\|_2^2}, \qquad \|w^*\|_2^2 = \frac{1}{\|Z^{\perp\{X,U\}}\|_2^2}.$$

By definition,

$$R_w^2 = 1 - \frac{\|w\|_2^2}{\|w^*\|_2^2} = 1 - \frac{\|Z^{\perp\{X,U\}}\|_2^2}{\|Z^{\perp X}\|_2^2}.$$

On the other hand, the partial R^2 from regressing Z on U conditional on X is

$$R_{Z\sim U|X}^2 = 1 - \frac{\|Z^{\perp\{X,U\}}\|_2^2}{\|Z^{\perp X}\|_2^2},$$

which is exactly the same quantity. Hence

$$R_{Z\sim U|X}^2 = R_w^2.$$

□

B.4.4 Proof of Theorem 7

Linear Regression (ATE)

Proof. The regression estimator of ATE from LR can be expressed as

$$\hat{\tau} = \sum_{i:Z_i=1} w_i Y_i - \sum_{i:Z_i=0} w_i Y_i$$

where

$$w_i = \begin{cases} 1/n_t + n/n_c (X_i - \bar{X}_t)^\top (S_t + S_c)^{-1} (\bar{X} - \bar{X}_t), & i \in \{i : Z_i = 1\} \\ 1/n_c + n/n_t (X_i - \bar{X}_c)^\top (S_t + S_c)^{-1} (\bar{X} - \bar{X}_c), & i \in \{i : Z_i = 0\} \end{cases}$$

$$S_t = \sum_{i:Z_i=1} (X_i - \bar{X}_t)(X_i - \bar{X}_t)^\top, \quad S_c = \sum_{i:Z_i=0} (X_i - \bar{X}_c)(X_i - \bar{X}_c)^\top$$

To calculate the oracle weights, we need to include U in the regression. For treated units,

$$w_i^* = \frac{1}{n_t} + \frac{n}{n_c} \begin{pmatrix} X_i^\top - \bar{X}_t^\top & U_i - \bar{U}_t \end{pmatrix} (S_t^* + S_c^*)^{-1} \begin{pmatrix} \bar{X} - \bar{X}_t \\ \bar{U} - \bar{U}_t \end{pmatrix}$$

where

$$\begin{aligned} S_t^* &= \sum_{i:Z_i=1} \begin{pmatrix} X_i - \bar{X}_t \\ U_i - \bar{U}_t \end{pmatrix} \begin{pmatrix} X_i^\top - \bar{X}_t^\top & U_i - \bar{U}_t \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i:Z_i=1} (X_i - \bar{X}_t)(X_i - \bar{X}_t)^\top & \sum_{i:Z_i=1} (X_i - \bar{X}_t)(U_i - \bar{U}_t) \\ \sum_{i:Z_i=1} (U_i - \bar{U}_t)(X_i - \bar{X}_t)^\top & \sum_{i:Z_i=1} (U_i - \bar{U}_t)^2 \end{pmatrix} \\ S_c^* &= \begin{pmatrix} \sum_{i:Z_i=0} (X_i - \bar{X}_c)(X_i - \bar{X}_c)^\top & \sum_{i:Z_i=0} (X_i - \bar{X}_c)(U_i - \bar{U}_c) \\ \sum_{i:Z_i=0} (U_i - \bar{U}_c)(X_i - \bar{X}_c)^\top & \sum_{i:Z_i=0} (U_i - \bar{U}_c)^2 \end{pmatrix} \end{aligned}$$

To simplify the notation, let

$$\begin{aligned} A &:= \sum_{i:Z_i=1} (X_i - \bar{X}_t)(X_i - \bar{X}_t)^\top + \sum_{i:Z_i=0} (X_i - \bar{X}_c)(X_i - \bar{X}_c)^\top = (X^{\perp Z})^\top X^{\perp Z} \\ b &:= \sum_{i:Z_i=1} (X_i - \bar{X}_t)(U_i - \bar{U}_t) + \sum_{i:Z_i=0} (X_i - \bar{X}_c)(U_i - \bar{U}_c) = (X^{\perp Z})^\top U^{\perp Z} \\ d &:= \sum_{i:Z_i=1} (U_i - \bar{U}_t)^2 + \sum_{i:Z_i=0} (U_i - \bar{U}_c)^2 = (U^{\perp Z})^\top U^{\perp Z} \end{aligned}$$

Invert the block matrix,

$$\begin{aligned} (S_t^* + S_c^*)^{-1} &= \begin{pmatrix} A & b \\ b^\top & d \end{pmatrix}^{-1} \\ &= \begin{pmatrix} A^{-1} + A^{-1} b (d - b^\top A^{-1} b)^{-1} b^\top A^{-1} & -A^{-1} b (d - b^\top A^{-1} b)^{-1} \\ -(d - b^\top A^{-1} b)^{-1} b^\top A^{-1} & (d - b^\top A^{-1} b)^{-1} \end{pmatrix} \end{aligned}$$

Then

$$
\begin{aligned}
w_i^* &= \frac{1}{n_t} + \frac{n}{n_c}\begin{pmatrix} X_i^\top - \bar{X}_t^\top & U_i - \bar{U}_t \end{pmatrix}\begin{pmatrix} A^{-1} + A^{-1}b(d - b^\top A^{-1}b)^{-1}b^\top A^{-1} & -A^{-1}b(d - b^\top A^{-1}b)^{-1} \\ -(d - b^\top A^{-1}b)^{-1}b^\top A^{-1} & (d - b^\top A^{-1}b)^{-1} \end{pmatrix}\begin{pmatrix} \bar{X} - \bar{X}_t \\ \bar{U} - \bar{U}_t \end{pmatrix} \\
&= \frac{1}{n_t} + \frac{n}{n_c}\Big[(X_i^\top - \bar{X}_t^\top)A^{-1}(\bar{X} - \bar{X}_t) + (U_i - \bar{U}_t)(d - b^\top A^{-1}b)^{-1}(\bar{U} - \bar{U}_t) \\
&\qquad - (U_i - \bar{U}_t)(d - b^\top A^{-1}b)^{-1}b^\top A^{-1}(\bar{X} - \bar{X}_t) - (X_i^\top - \bar{X}_t^\top)A^{-1}b(d - b^\top A^{-1}b)^{-1}(\bar{U} - \bar{U}_t) \\
&\qquad + (X_i^\top - \bar{X}_t^\top)A^{-1}b(d - b^\top A^{-1}b)^{-1}b^\top A^{-1}(\bar{X} - \bar{X}_t)\Big] \\
&= \frac{1}{n_t} + \frac{n}{n_c}(X_i^\top - \bar{X}_t^\top)A^{-1}(\bar{X} - \bar{X}_t) + \frac{n}{n_c}(d - b^\top A^{-1}b)^{-1}\Big[(U_i - \bar{U}_t)(\bar{U} - \bar{U}_t) - (U_i - \bar{U}_t)b^\top A^{-1}(\bar{X} - \bar{X}_t) \\
&\qquad - (X_i^\top - \bar{X}_t^\top)A^{-1}b(\bar{U} - \bar{U}_t) + (X_i^\top - \bar{X}_t^\top)A^{-1}bb^\top A^{-1}(\bar{X} - \bar{X}_t)\Big] \\
&= w_i + \frac{n}{n_c}(d - b^\top A^{-1}b)^{-1} \\
&\qquad \left[(U_i - \bar{U}_t) - (X_i^\top - \bar{X}_t^\top)A^{-1}b\right]\left[(\bar{U} - \bar{U}_t) - b^\top A^{-1}(\bar{X} - \bar{X}_t)\right]
\end{aligned}
$$

Notice that

$$
\begin{gathered}
A^{-1}b = [(X^{\perp Z})^\top X^{\perp Z}]^{-1}(X^{\perp Z})^\top U^{\perp Z} = \hat{\beta}_{U^{\perp z} \sim X^{\perp z}} = \hat{\beta}_{U \sim X|Z} \\
d - b^\top A^{-1}b = (U^{\perp Z})^\top \left(I - P_{X^{\perp z}}\right) U^{\perp Z}
\end{gathered}
$$

Therefore, we get

$$
w_i^* - w_i = m\left[U_i - \bar{U}_t - (X_i - \bar{X}_t)^\top v\right]\left[\bar{U}_c - \bar{U}_t - (\bar{X}_c - \bar{X}_t)^\top v\right]
$$

where

$$
v = \hat{\beta}_{U \sim X|Z} \text{ and } m = \frac{1}{(U^{\perp Z})^\top \left(I - P_{X^{\perp z}}\right) U^{\perp Z}} = 1/(n \operatorname{Var}(U - X^\top v|Z))
$$

And the proof for controlled units follow exactly the same logic.

□

Interacted Linear Regression (ATT)

Proof. The regression estimator of ATT from ILR can be expressed as

$$
\hat{\tau} = \sum_{i:Z_i=1} w_i Y_i - \sum_{i:Z_i=0} w_i Y_i
$$

where

$$
w_i = \begin{cases} 1/n_t & i \in \{i : Z_i = 1\} \\ 1/n_c + (X_i - \bar{X}_c)^\top S_c^{-1}(\bar{X}_t - \bar{X}_c), & i \in \{i : Z_i = 0\} \end{cases}
$$

We only need to consider the control units here. Including the unobserved confounder U and following the same matrix calculation as in the LR case, we have

$$w_i^* = w_i + (d - b^\top A^{-1} b)^{-1} \left[(U_i - \bar{U}_c) - (X_i^\top - \bar{X}_c^\top) A^{-1} b \right] \left[(\bar{U}_t - \bar{U}_c) - b^\top A^{-1} (\bar{X}_t - \bar{X}_c) \right]$$

where

$$\begin{aligned} A &:= \sum_{i:Z_i=0} (X_i - \bar{X}_c)(X_i - \bar{X}_c)^\top = X_c^\top X_c \\ b &:= \sum_{i:Z_i=0} (X_i - \bar{X}_c)(U_i - \bar{U}_c) = X_c^\top U_c \\ d &:= \sum_{i:Z_i=0} (U_i - \bar{U}_c)^2 = U_c^\top U_c \end{aligned}$$

Then we can write

$$w_i^* - w_i = -m \left[U_i - \bar{U}_c - (X_i - \bar{X}_c)^\top v \right] \left[\bar{U}_c - \bar{U}_t - (\bar{X}_c - \bar{X}_t)^\top v \right]$$

where

$$\begin{aligned} v &= A^{-1} b = (X_c^\top X_c)^{-1} X_c^\top U_c = \hat{\beta}_{U_c \sim X_c} \\ m &= (d - b^\top A^{-1} b)^{-1} = (U_c^\top U_c - U_c^\top X_c (X_c^\top X_c)^{-1} X_c^\top U_c)^{-1} = \frac{1}{U_c^\top (I - P_{X_c}) U_c} \end{aligned}$$

□

Interacted Linear Regression (ATE)

Proof. The regression estimator of ATE from ILR can be expressed as

$$\hat{\tau} = \sum_{i:Z_i=1} w_i Y_i - \sum_{i:Z_i=0} w_i Y_i$$

where

$$w_i = \begin{cases} 1/n_t + (X_i - \bar{X}_t)^\top S_t^{-1} (\bar{X} - \bar{X}_t), & i \in \{i : Z_i = 1\} \\ 1/n_c + (X_i - \bar{X}_c)^\top S_c^{-1} (\bar{X} - \bar{X}_c), & i \in \{i : Z_i = 0\} \end{cases}$$

Including the unobserved confounder U here, the proof is almost the same as the ATT case, and we just need to follow the same logic for the treated group as well. □

Appendix C

Supplementary Material for Chapter 3

This supplement contains discussion of an alternate effect estimator based on inverse weighting, and proofs of two results from the main manuscript: an asymptotic guarantee of type I error control for covariate-adaptive randomization inference with estimated propensity scores when treatment follows a logistic model and a guarantee that the sensitivity analysis method described in the main manuscript controls Type I error under limited unobserved confounding (given true propensity scores). In addition, an alternative large-sample variance estimator for the null distribution of the mean-corrected difference-in-means estimator under the covariate-adaptive randomization distribution is constructed, and supporting results for the simulation study in the main manuscript are provided.

C.1 Inverse propensity weighting estimator

One frustrating aspect of the difference-in-means statistic discussed in Section 3.2 in the context of covariate-adaptive randomization inference is that it is no longer unbiased for the constant additive effect of treatment τ in model (4) (as it would be in a randomized trial or a study with exact matching). While the maximum p-value estimator described in Section 3.3 enjoys many good properties, its expectation in finite samples is difficult to characterize. Instead, consider the following estimator, based on inverse probability weighting [59]. For mathematical convenience, we focus on the case of matched pairs and assume $n_k = 2$ throughout this section:

$$\begin{aligned} T_{IPW}(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}) &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\left(\sum_{i=1}^{2}\frac{Z_{ki}Y_{ki}}{p_{ki}} - \sum_{i=1}^{2}\frac{(1-Z_{ki})Y_{ki}}{1-p_{ki}}\right) \\ &= \frac{1}{2K}\sum_{k=1}^{K}(Y_{k1}-Y_{k2})\frac{Z_{k1}-p_{k1}}{p_{k1}(1-p_{k1})} \end{aligned}$$

where p_{ki} is defined in Section 3.1 in the main manuscript. When true propensity scores are used to calculate the p_{ki} terms in the denominator, this estimator is unbiased for τ

under the constant additive model (equation (4) in the main manuscript) under treatment distributed as in equation (2) in the main manuscript. Covariate-adaptive randomization tests and associated confidence intervals could be conducted using this test statistic instead of the difference-in-means. Some aspects of the process such as inverting the confidence interval might become more difficult—for example, the shortcut formula for quickly inverting test statistics in Section 3.3 would become more complicated—while some aspects would be easier; in particular, the test statistic itself could be used as a point estimate. The sensitivity analysis results of Section 5 apply to any sum statistic $T(\mathbf{Z}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} Z_{ki} f_{ki}$ where f_{ki} is some function of $\mathbf{Y}$; the inverse weighting estimator qualifies (with $f_{ki} = (Y_{ki} - Y_{k(3-i)})/p_{ki}$), so sensitivity analysis works for this estimator as well.

To get more insight into the pros and cons of the two estimators we compare their expectations and variances when a constant additive effect τ is present:

$$
\begin{aligned}
E(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})) &= \tau + \frac{1}{K} \sum_{k=1}^{K} (p_{k1} - p_{k2})(Y_{k1}(0) - Y_{k2}(0)) \\
E(T_{IPW}(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})) &= \tau \\
Var(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})) &= \frac{1}{K^2} \sum_{k=1}^{K} 4p_{k1}(1 - p_{k1})(Y_{k1}(0) - Y_{k2}(0))^2 \\
Var(T_{IPW}(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}})) &= \frac{1}{K^2} \sum_{k=1}^{K} \frac{[(Y_{k1}(0) - Y_{k2}(0)) + (1 - 2p_{k1})\tau]^2}{4p_{k1}(1 - p_{k1})}.
\end{aligned}
$$

While T_{IPW} is clearly superior to the difference in means in terms of bias, it may also exhibit a much larger variance. Notice that for any $p \in (0, 1)$, $4p(1-p) \leq 1 \leq 1/(4p(1-p))$, indicating that when $\tau = 0$, $Var(T(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}))$ is always smaller than $Var(T_{IPW}(\mathbf{Z}_{\mathcal{M}}, \mathbf{Y}_{\mathcal{M}}))$ unless $p_{k1} = 1/2$ or $Y_{k1} = Y_{k2}$ for any k. In addition, when τ is sufficiently large in magnitude relative to the quantities $Y_{k1}(0) - Y_{k2}(0)$ the inverse weighted estimator will also exhibit a larger variance.

As in Section 6.3 of the main manuscript ($n = 100$, $p = 2$, strong propensity signal, propensity scores estimated in-sample[1]), we set the true treatment effect τ to each of four values: 0, 1, 10, and 100. Table C.1 compares the performance of the maximum p-value estimator and the IPW estimator under these conditions. In scenarios without Z-dependence, estimates calculated using the maximum p-value and IPW strategies appear approximately unbiased while DiM estimates are biased. The maximum p-value estimates tend to be less variable than the IPW estimates, and in cases where IPW performs better, the differences are small. In addition, as shown in the above variance calculation and simulation with $\tau = 100$, the variance of IPW estimator explodes with very large τ while the variance of the maximum p-value and DiM estimator do not change with τ. When Z-dependence is present, all three

[1]Section 6.3's simulation setup turned out to be a more interesting case than Section 6.2's, which is similar but with a weaker outcome signal. Initial experiments with the latter setup revealed almost identical performance between the estimators.

estimators will have bigger biases but the maximum p-value still performs better than IPW in general, and both are much better than DiM.

C.2 Proof of Theorem 2

Proof. We use Pinsker's inequality to bound the quantity $d_{TV}(P_\lambda, P_{\widehat{\lambda}})$ by the square root of the sum of the Kullback-Leibler divergences between true and estimated conditional distributions of treatment for each unit in the study, then use a Taylor expansion to show that this quantity converges in probability to zero. In combination with Theorem 1,this suffices for the result.

We let $\widehat{\beta}$ represent the maximum likelihood estimate of β obtained from the size-N pilot sample.

$$\begin{aligned}
d_{TV}^2(P_\lambda, P_{\widehat{\lambda}}) &\leq \frac{1}{2}\sum_{i=1}^{n} d_{KL}\left(P_{\lambda_i}, P_{\widehat{\lambda}_i}\right) \\
&= \frac{1}{2}\sum_{i=1}^{n} E\log\left\{\left(\frac{1+\exp[-\widehat{\beta}^T x_i]}{1+\exp[-\beta^T x_i]}\right)^{Z_i}\left(\frac{1+\exp[\widehat{\beta}^T x_i]}{1+\exp[\beta^T x_i]}\right)^{1-Z_i}\right\} \\
&= \frac{1}{2}\sum_{i=1}^{n} E\left\{Z_i\log\left(\frac{1+\exp[-\widehat{\beta}^T x_i]}{1+\exp[-\beta^T x_i]}\right) + (1-Z_i)\log\left(\frac{1+\exp[\widehat{\beta}^T x_i]}{1+\exp[\beta^T x_i]}\right)\right\} \\
&= \frac{1}{2}\sum_{i=1}^{n} E\left\{\frac{1}{1+\exp[-\beta^T x_i]}\log\left(\frac{1+\exp[-\widehat{\beta}^T x_i]}{1+\exp[-\beta^T x_i]}\right) + \frac{1}{1+\exp[\beta^T x_i]}\log\left(\frac{1+\exp[\widehat{\beta}^T x_i]}{1+\exp[\beta^T x_i]}\right)\right\} \\
&= \frac{1}{2}\sum_{i=1}^{n}\left\{C_{1i}f(-\widehat{\beta}^T x_i) + C_{2i}f(\widehat{\beta}^T x_i) + C_{3i}\right\} = g(\widehat{\beta}).
\end{aligned}$$

where

$$\begin{aligned}
C_{1i} &= 1/(1+\exp[-\beta^T x_i]) \in (0,1) \\
C_{2i} &= 1/(1+\exp[\beta^T x_i]) \in (0,1) \\
C_{3i} &= -C_{1i}\log(1+\exp[-\beta^T x_i]) - C_{2i}\log(1+\exp[\beta^T x_i]) \\
f(b) &= \log(1+\exp(b))
\end{aligned}$$

We use a first-order multivariate Taylor expansion to understand the behavior of the vector-valued nonlinear function g around the vector $\beta^T X$. First note that

$$\begin{aligned}
f'(b) &= \frac{\exp(b)}{1+\exp(b)} = \frac{1}{1+\exp(-b)} \\
f''(b) &= \frac{\exp(-b)}{(1+\exp(-b))^2} = \frac{1}{\exp(b)+2+\exp(-b)} \leq \frac{1}{2}
\end{aligned}$$

so:

$$\nabla g(\beta^T X) = \frac{1}{2}\sum_{i=1}^{n}\left\{-C_{1i}f'(-\beta^T x_i) + C_{2i}f'(\beta^T x_i)\right\}x_i = \frac{1}{2}\sum_{i=1}^{n}\left\{-C_{1i}C_{2i} + C_{2i}C_{i1}\right\}x_i = \mathbf{0}.$$

$$H_g(\beta^T X) = \frac{1}{2}\sum_{i=1}^{n}\left\{C_{1i}f''(-\beta^T x_i) + C_{2i}f''(\beta^T x_i)\right\}x_i x_i^T$$

Taylor's theorem gives us the following:

$$\begin{aligned} g(\widehat{\beta}) &= g(\beta) + \nabla g(\beta^T X)^T(\widehat{\beta} - \beta) + R_2 \\ &= 0 + 0 + R_2 \end{aligned}$$

Furthermore, using the Lagrange bound we can bound the second-order remainder R_2 (letting $|A|$ represent a matrix with each of its elements replaced by its absolute value). Specifically, for some $\widetilde{\beta}$ near β, we have the following:

$$\begin{aligned} g(\widehat{\beta}) &\leq (\widehat{\beta} - \beta)^T |H_g(\widetilde{\beta})|(\widehat{\beta} - \beta) \\ &\leq \frac{1}{2}\sum_{i=1}^{n}\left|C_{1i}f''(-\beta^T x_i) + C_{2i}f''(\beta^T x_i)\right|(\widehat{\beta} - \beta)^T |x_i x_i^T|(\widehat{\beta} - \beta) \\ &\leq \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{2} + \frac{1}{2}\right)(\widehat{\beta} - \beta)^T |x_i x_i^T|(\widehat{\beta} - \beta) \\ &= \frac{1}{2}\sum_{i=1}^{n}\frac{\left[\sqrt{N}(\widehat{\beta} - \beta)^T |x_i|\right]^2}{N} \end{aligned} \tag{C.1}$$

Since $\widehat{\beta}$ is a maximum likelihood estimate fitted on a sample of size N, for any x_i the quantity $\sqrt{N}(\widehat{\beta} - \beta)^T x_i$ converges in distribution to a normal random variable with mean zero and variance $x_i^T \Sigma x_i$ for some variance-covariance matrix Σ, and by the continuous mapping theorem, each term in summation (C.1) converges in distribution to

$$\frac{W x_i^T \Sigma x_i}{N}$$

where $W \sim \chi_1^2$. Since x_i is chosen from compact support, there is a uniform upper bound $\widetilde{\sigma}^2$ on these variances $x_i^T \Sigma x_i$ for all $i = 1, \ldots, n$. Then for any $\epsilon > 0$ and any $\delta > 0$, we may choose $N_{\epsilon,\delta}$ sufficiently large to ensure that for all $N \geq N_{\epsilon,\delta}$,

$$P\left(W > \left(\frac{N}{\widetilde{\sigma}^2}\right)\delta\right) < \epsilon$$

Thus each term in summation (C.1) is $O_p\left(\frac{\widetilde{\sigma}^2}{N}\right)$, and in turn the entire summation is $O_p\left(\frac{n\widetilde{\sigma}^2}{N}\right)$, and it square root is $O_p\left(\sqrt{\frac{n\widetilde{\sigma}^2}{N}}\right)$. Since $\lim_{n\longrightarrow\infty} n/N = 0$, this is sufficient for $\sqrt{g(\widehat{\beta})}$ to converge in probability to zero.

To complete the proof, fix any $\epsilon > 0$. We now establish that there exists N_ϵ such that for all $N > N_\epsilon$,

$$P(p_{\widehat{\lambda}_N,n} \leq \alpha) - \alpha < \epsilon.$$

which suffices for the desired result. To do this, note that the convergence of $d_{TV}(P_\lambda, P_{\widehat{\lambda}})$ to zero in probability implies that there exists N' such that for all $N > N'$,

$$P(d_{TV}(P_\lambda, P_{\widehat{\lambda}}) > \epsilon/2) < \epsilon/2.$$

Then when $N > N'$, by Theorem 1:

$$\begin{aligned} P(p_{\widehat{\lambda}_N,n} \leq \alpha) - \alpha &\leq d_{TV}(P_\lambda, P_{\widehat{\lambda}}) \\ &= 1\{d_{TV}(P_\lambda, P_{\widehat{\lambda}}) \leq \epsilon/2\} d_{TV}(P_\lambda, P_{\widehat{\lambda}}) + 1\{d_{TV}(P_\lambda, P_{\widehat{\lambda}}) > \epsilon/2\} d_{TV}(P_\lambda, P_{\widehat{\lambda}}) \end{aligned}$$

so

$$\begin{aligned} E_{\widehat{\lambda}}\left[P(p_{\widehat{\lambda}_N,n} \leq \alpha) - \alpha\right] &\leq E_{\widehat{\lambda}}\left[1\{d_{TV}(P_\lambda, P_{\widehat{\lambda}}) \leq \epsilon/2\} d_{TV}(P_\lambda, P_{\widehat{\lambda}})\right] + E_{\widehat{\lambda}}\left[1\{d_{TV}(P_\lambda, P_{\widehat{\lambda}}) > \epsilon/2\} d_{TV}(P_\lambda, P_{\widehat{\lambda}})\right] \\ P(p_{\widehat{\lambda}_N,n} \leq \alpha) - \alpha &\leq \epsilon/2 + P_{\widehat{\lambda}}(d_{TV}(P_\lambda, P_{\widehat{\lambda}}) > \epsilon/2) \\ &< \epsilon/2 + \epsilon/2 = \epsilon \end{aligned}$$

as desired. As indicated, expections are taken with respect to the sampling distribution of the estimated propensity scores, which does not affect the left-hand side because it is a constant. The second-to-last line follows from the law of total probability and the upper bound of 1 on the total variation distance. □

C.3 Proof of Theorem 3

Proof. Assuming model (6) from the main manuscript for treatment assignment and test statistic $\sum_{k=1}^{K}\sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y}) = \mathbf{Z}^T\mathbf{f}$, define:

$$\omega(\mathbf{f}, \mathbf{u}) = E_{\mathbf{Z}}\{h(\mathbf{Z}, \mathbf{f})\} = \sum_{\mathbf{z}\in\boldsymbol{\Omega}} h(\mathbf{z}, \mathbf{f}) P(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{z}\in\boldsymbol{\Omega}} h(\mathbf{z}, \mathbf{f}) \prod_{k=1}^{K} \frac{\exp\left\{\mathbf{z_k}^T(\boldsymbol{\kappa}_k + \gamma \mathbf{u}_k)\right\}}{\sum_{\mathbf{b}\in\boldsymbol{\Omega}_k} \exp\left\{\mathbf{b}^T(\boldsymbol{\kappa_k} + \gamma \mathbf{u_k})\right\}}$$

where $h(\mathbf{z}, \mathbf{f}) = \mathbb{1}\left\{\mathbf{z}^T\mathbf{f} \geq C\right\}$. By choosing C equal to the critical value for our one-sided test, we obtain $\omega(\mathbf{f}, \mathbf{u}) = \alpha_{adapt}(\mathbf{Y}, \mathbf{u})$, so it will be sufficient to show that

$$\min_{\mathbf{u}'\in U^-} \omega(\mathbf{f}, \mathbf{u}') \leq \omega(\mathbf{f}, \mathbf{u}) \leq \max_{\mathbf{u}\in U^+} \omega(\mathbf{f}, \mathbf{u}). \tag{C.2}$$

To establish (C.2), we make two separate sub-arguments. First, we prove the following lemma analogous to Proposition 18 of Rosenbaum [77], which suffices to show that extrema are achieved only when the vector $\mathbf{u}$ is binary. Secondly, we argue that at the maximizing (minimizing) value of $\mathbf{u}$, in any matched set k, $f_{ki} < f_{kj}$ implies $u_{ki} \leq u_{kj}$ ($u_{ki} \geq u_{kj}$).

Lemma 1. *Let $\mathbf{e}_{\ell i}$ be a vector in $\mathbb{R}^n$ containing a one in index i of matched set ℓ and zeroes in all other indices. For each fixed $\mathbf{f}, \mathbf{u}$, and i, $\omega(\mathbf{f}, \mathbf{u} + \delta\mathbf{e}_{\ell i})$ is monotone in δ.*

Proof. Let $\mathbf{\Omega}_1$ be the set of vectors $\mathbf{b} \in \mathbf{\Omega}$ for which $b_{\ell i} = 1$, and let $\mathbf{\Omega}_0 = \mathbf{\Omega} - \mathbf{\Omega}_1$.

$$\begin{aligned}
\omega(\mathbf{f}, \mathbf{u} + \delta\mathbf{e}_{\ell i}) &= \frac{\sum_{\mathbf{z}\in\mathbf{\Omega}} h(\mathbf{z}, \mathbf{f}) \exp\left\{\mathbf{z}^T(\boldsymbol{\kappa} + \gamma\mathbf{u} + \gamma\delta\mathbf{e}_{\ell i})\right\}}{\sum_{\mathbf{b}\in\mathbf{\Omega}} \exp\left\{\mathbf{b}^T(\boldsymbol{\kappa} + \gamma\mathbf{u} + \gamma\delta\mathbf{e}_{\ell i})\right\}} \\
&= \frac{\sum_{\mathbf{z}\in\mathbf{\Omega}_0} h(\mathbf{z}, \mathbf{f}) \exp\left\{\mathbf{z}^T(\boldsymbol{\kappa} + \gamma\mathbf{u})\right\} + \sum_{\mathbf{z}\in\mathbf{\Omega}_1} h(\mathbf{z}, \mathbf{f}) \exp\left\{\mathbf{z}^T(\boldsymbol{\kappa} + \gamma\mathbf{u}) + \gamma\delta\right\}}{\sum_{\mathbf{b}\in\mathbf{\Omega}_0} \exp\left\{\mathbf{b}^T(\boldsymbol{\kappa} + \gamma\mathbf{u})\right\} + \sum_{\mathbf{b}\in\mathbf{\Omega}_1} \exp\left\{\mathbf{b}^T(\boldsymbol{\kappa} + \gamma\mathbf{u}) + \gamma\delta\right\}} \\
&= \frac{A_0 + \exp(\delta\gamma)A_1}{D_0 + \exp(\delta\gamma)D_1}
\end{aligned}$$

where A_i, D_i do not depend on δ and the D_i are all strictly positive. From this point on the proof is identical to that of Proposition 18 in Rosenbaum [77]: the partial derivative with respect to δ (computed below) has constant sign, which is sufficient for monotonicity.

$$\frac{\partial\omega(\mathbf{f}, \mathbf{u} + \delta\mathbf{e}_{\ell i})}{\partial\delta} = \frac{(D_0A_1 - A_0D_1)\gamma_\ell \exp(\gamma\delta)}{\left[D_0 + D_1 \exp(\gamma\delta)\right]^2}$$

□

Lemma 1 tells us that if there exists some $\mathbf{u}$-element $u_{ki} \in (0, 1)$, then either increasing u_{ki} to one or decreasing it to zero will increase (or at least not decrease) the associated value of $\omega(\mathbf{f}, \mathbf{u})$; similarly, shifting u_{ki} to the opposite extreme will decrease (or at least not increase) $\omega(\mathbf{f}, \mathbf{u})$. As such we can restrict attention to binary $\mathbf{u}$.

Now consider the ordering of the elements of $\mathbf{u}$ at the extrema. We present the proof for the maximum only, since the case of the minimum proceeds by near-identical arguments. It suffices to show that when $h(\mathbf{z}, \mathbf{f}) = g(\mathbf{z}, \mathbf{f})$, the solution $\mathbf{u}$ to $\max_{\mathbf{u}\in[0,1]^n} \omega(\mathbf{f}, \mathbf{u})$ satisfies $u_{ki} \leq u_{kj}$ whenever $f_{ki} < f_{kj}$.

Suppose not. Then there exists at least one stratum k with indices j, j' such that $u_{kj} > u_{kj'}$ but $f_{kj} < f_{kj'}$. We expand $\omega(\mathbf{f}, \mathbf{u})$ as follows. We index by k $(-k)$ to select subvectors of vectors in $\mathbb{R}^n$ containing only elements in (not in) stratum k, and we let Ω_k and Ω_{-k} represent the set of all possible values for $\mathbf{z}_k$ and $\mathbf{z}_{-k}$.

$$\begin{aligned}
\omega(\mathbf{f}, \mathbf{u}) &= \sum_{\mathbf{z}\in\Omega} h(\mathbf{z}, \mathbf{f}) P(\mathbf{Z} = \mathbf{z}) \\
&= \sum_{\mathbf{z}_k\in\Omega_k} P(\mathbf{Z}_k = \mathbf{z}_k) \sum_{\mathbf{z}_{-k}\in\Omega_{-k}} h(\mathbf{z}, \mathbf{f}) P(\mathbf{Z}_{-k} = \mathbf{z}_{-k}) \\
&= \sum_{\mathbf{z}_k\in\Omega_k} P(\mathbf{Z}_k = \mathbf{z}_k) E\left[\mathbb{1}\left\{\mathbf{z}_{-k}^T\mathbf{f}_{-k} + \mathbf{z}_k^T\mathbf{f}_k > C\right\} \middle| \mathbf{Z}_k = \mathbf{z}_k\right]
\end{aligned}$$

Let $h^*(\mathbf{z}_k, \mathbf{f}_k) = E\left[\mathbb{1}\left\{\mathbf{z}_{-k}^T\mathbf{f}_{-k} + \mathbf{z}_k^T\mathbf{f}_k > C\right\} \middle| \mathbf{Z}_k = \mathbf{z}_k\right]$. Holding $\mathbf{u}_{-k}$ fixed, we can rewrite the problem of maximizing $\omega(f, u)$ over $\mathbf{u}_k$ as follows (abusing notation slightly to let $\mathbf{e}_{ki}$ indicate the vector in $\mathbb{R}^{n_k}$ with zeroes at all indices except for a one at index i):

$$\max \frac{\sum_{i=1}^{n_k} h^*(\mathbf{e}_{ki}, \mathbf{f}_k) \exp\left\{\kappa_{ki} + \gamma u_{ki}\right\}}{\sum_{i=1}^{n_k} \exp\left\{\kappa_{ki} + \gamma u_{ki}\right\}} \qquad \text{s.t. } \mathbf{u}_k \in [0, 1]^{n_k}$$

Following a similar argument of Zhao, Small, and Bhattacharya [101], we now use the Charnes-Cooper transform to write it as a linear program by defining $w_{ki} = \frac{\exp(\gamma u_{ki})}{\sum_{i=1}^{n_k} \exp\{\kappa_{ki} + \gamma u_{ki}\}}$ and $t = \frac{1}{\sum_{i=1}^{n_k} \exp\{\kappa_{ki} + \gamma u_{ki}\}}$, which gives the following equivalent form:

$$\begin{aligned}
&\max \sum_{i=1}^{n_k} h^*(\mathbf{e}_{ki}, \mathbf{f}_k) \exp(\kappa_{ki}) w_{ki} \\
&\text{s.t.} \\
&t \leq w_{ki} \leq \exp(\gamma) t \\
&\sum_{i=1}^{n_k} \exp(\kappa_{ki}) w_{ki} = 1 \\
&t \geq 0.
\end{aligned}$$

Let $(\mathbf{w}_k, t)$ be the optimal solution derived from $\mathbf{u}_k$. The ordering of the w_{ki}s is identical to the ordering of the u_{ki}s, so that $w_{kj} > w_{kj'}$; in addition, note that $f_{kj} < f_{kj'}$ implies $h^*(\mathbf{e}_{kj}, \mathbf{f}_k) < h^*(\mathbf{e}_{kj'}, \mathbf{f}_k)$ by construction of h^*. For sufficiently small $\epsilon > 0$, the solution $(\mathbf{w}^*, t^*)$ is also feasible where $t^* = t$ and

$$w^*_{ki} = \begin{cases} w_{kj} - \epsilon \exp(-\kappa_{kj}) & \text{if } i = j \\ w_{kj'} + \epsilon \exp(-\kappa_{kj'}) & \text{if } i = j' \\ w_{ki} & \text{otherwise} \end{cases}$$

The objective value under $(\mathbf{w}^*_k, t^*)$ is equal to the objective value under $(\mathbf{w}_k, t)$ plus the quantity $\epsilon[h^*(\mathbf{e}_{kj'}, \mathbf{f}_k) - h^*(\mathbf{e}_{kj}, \mathbf{f}_k)]$ which is positive under our assumptions, but this means $(\mathbf{w}_k, t)$ cannot be optimal. Therefore in any optimal solution, $f_{kj} < f_{kj'}$ implies $u_{kj} \leq u_{kj'}$ (or $w_{kj} \leq w_{kj'}$) and the proof of the theorem is complete. □

C.4 Alternative large-sample null distribution accounting for propensity score estimation

In this section we construct and give intuition for an alternative variance estimator for the difference between the difference-in-means statistic and its estimated mean under covariate-adaptive randomization inference that accounts for estimation of the propensity score. Specifically, we are motivated by the large-sample distribution derived for the difference-in-means statistic in Section 3.2 of the main manuscript, which gives the following form for the mean of the difference-in-means estimator:

$$\mu_0 = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k p_{ki} - 1}{n_k - 1}.$$

Section 3.2 suggests a large-sample version of the covariate-adaptive randomization test that compares the difference-in-means statistic T to a normal distribution centered at μ_0, or

equivalently compares $T - \mu_0$ to a normal distribution centered at zero. However, since μ_0 depends on the unknown propensity score it must be estimated, typically using the equation above but plugging in propensity score estimates for p_{ki} to give $\widehat{\mu}$. This leads the variance of $T - \widehat{\mu}$ to differ from the variance of T or $T - \mu_0$. To construct an estimate of the variance of $\widetilde{T} = T - \widehat{\mu}$ we represent both propensity score estimation and calculation of $\widetilde{T}$ as tasks within a common M-estimation framework.

While M-estimators are typically used in settings where independent samples from an infinite superpopulation are available and the goal is to estimate a parameter of that infinite population [91], we focus instead on a setting with an infinite sequence of ever-larger finite populations, in each of which only the treatment random variable $\mathbf{Z}$ is random. In addition, some subjects within each finite population are grouped into matched sets, and treatment indicators for subjects within the same matched set are highly dependent; on the other hand, propensity score estimation typically leverages all unmatched individuals as well as those in matched sets, and it is more reasonable to think of unmatched individuals' treatments as mutually independent. As such, we adopt new notation as follows.

First, for a given finite population, we let $\mathbf{O}_k = \{\mathbf{Y}_k, \mathbf{Z}_k, \mathbf{X}_k\}$ represent the set of observed data associated with all the units in matched sets $k = 1, \ldots K$. We also introduce quantities $\mathbf{O}_{K+1}, \ldots, \mathbf{O}_{K'}$ where each $\mathbf{O}_k$ with $k > K$ gives the Y, Z, X information associated with a single unmatched unit, so $n_k = 1$ for these units. Letting $m = \sum_{k=1}^{K'} n_k$, we define $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{Y}(z)$ as M-tuples collecting values from all K' sets, with $\mathbf{X}$ defined as the analogous $m \times p$ matrix. We let $\mathbf{Z}_k$ be independent of $\mathbf{Z}_{k'}$ for all $1 \leq k, k' \leq K', k \neq k'$, conditional on $\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X}$.

We assume that the propensity score is fit using M-estimation (such as a logistic regression). Let $\theta \in \mathbb{R}^p$ be the parameter vector for the propensity score model, and let

$$\boldsymbol{\psi}(\mathbf{O}_k, \theta) = (\psi_1(\mathbf{O}_k, \theta), \ldots, \psi_p(\mathbf{O}_k, \theta))^T$$

be the set of associated functions that define the M-estimator via the estimating equations:

$$\sum_{k=1}^{K'} \boldsymbol{\psi}(\mathbf{O}_k, \theta) = \mathbf{0}.$$

Note that for convenience we assume matched sets are given here and will not be influenced by the estimated propensity score (see Section 2.3 of the main manuscript for more discussion of such assumptions and their implications). We augment these equations with three more components. The first, η, represents the inverse of the proportion of sets from 1 to K' that are included in the match. This is helpful since the propensity-score fitting parts of the common framework normalize over all observations, while most of the added components normalize only over matched sets; η rescales the latter estimating equations so they have the appropriate denominator. The other two added components are μ, the mean of the difference-in-means statistic, and our statistic of interest $\widetilde{T} = T - \widehat{\mu}$. μ is added as its own

entry in the parameter vector to make it easy to represent $\widetilde{T}$ in the M-estimation framework.

$$\sum_{k=1}^{K'} \begin{pmatrix} \boldsymbol{\psi}(\mathbf{O}_k, \theta) \\ 1/\eta - 1\{n_k > 1\} \\ \eta 1\{n_k > 1\} \left\{ \mu - \sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k \cdot \text{odds}_{ki}(\theta)}{(n_k - 1) \sum_{j=1}^{n_k} \text{odds}_{kj}(\theta)} - \frac{1}{n_k - 1} \right) \right\} \\ \eta 1\{n_k > 1\} \left\{ \widetilde{T} - \sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1} + \mu \right\} \end{pmatrix} = \mathbf{0}$$

We let $\boldsymbol{\psi}^{full}$ represent this expanded set of estimating equations. We may obtain parameter estimates $(\widehat{\theta}, \widehat{\mu}, \widetilde{T})$ by solving this equation. Note that our two definitions of $\widehat{\mu}$ agree by construction, and that $\widetilde{T}$ is simply the actual mean-corrected difference-in-means statistic rather than an estimate of it).

Assume that θ_0, μ_0, η_0 and $\widetilde{T}_0$ exist as unique solutions to the following equation:

$$\lim_{K' \longrightarrow \infty} \frac{1}{K'} \sum_{k=1}^{K'} E \left[\boldsymbol{\psi}^{full}(\mathbf{O}_k, \theta, \eta, \mu, \widetilde{T}) \mid \mathbf{Y}(1), \mathbf{Y}(0), \mathbf{X} \right].$$

In short, θ_0, μ_0, and η_0 are the large-sample limits of the propensity score parameter, the mean of the difference-in-means statistic under the covariate-adaptive randomization distribution, the ratio K'/K, and our chosen test statistic itself across our sequence of growing finite populations under the sharp null hypothesis. Note that under these conditions we know that our new and earlier definitions of μ_0 agree, and that $\widetilde{T}_0 = 0$.

We now wish to estimate the asymptotic variance of $\widetilde{T}$ under the sharp null hypothesis of no effect of treatment, accounting for the estimation of the propensity score. We construct an estimator using the sandwich variance approach described for traditional infinite-superpopulation M-estimators in Stefanski and Boos [91]. Specifically, we construct an estimator for one scalar entry of the following matrix, analogous to the (stabilized) asymptotic variance-covariance matrix of the parameter vector in the traditional infinite superpopulation version of M-estimation:

$$V(\theta_0, \eta_0, \mu_0) = A(\theta_0, \eta_0, \mu_0, 0)^{-1} B(\theta_0, \eta_0, \mu_0, 0) [A(\theta_0, \eta_0, \mu_0, 0)^{-1}]^T$$

where $A(\cdot)$ and $B(\cdot)$ are defined as follows:

$$A(\theta_0, \eta_0, \mu_0, 0) = \lim_{K' \longrightarrow \infty} \frac{1}{K'} \sum_{k=1}^{K'} E[-\nabla_{\theta, \eta, \mu, \widetilde{T}}, \boldsymbol{\psi}^{full}(\mathbf{O}_k, \theta_0, \eta_0, \mu_0, 0)]$$

$$B(\theta_0, \mu_0, 0) = \lim_{K' \longrightarrow \infty} \frac{1}{K'} \sum_{k=1}^{K'} E[\boldsymbol{\psi}^{full}(\mathbf{O}_k, \theta_0, \eta_0, \mu_0, 0) \boldsymbol{\psi}^{full}(\mathbf{O}_k, \theta_0, \eta_0, \mu_0, 0)^T]$$

Note that the formulas differ from the traditional ones in Stefanski and Boos [91], which take expectation over a sampling distribution for all observed elements O_i, by taking the

limit of a sample average over the non-random elements (Y_{ki}); we assume here that those limits exist and that the limit $A(\theta_0, \eta_0, \mu_0)$ is invertible. We acknowledge that our use of this formula in the finite population is heuristic in the sense that we do not provide formal regularity conditions or a result guaranteeing that the asymptotic variance of $\widetilde{T}$ is equal to $V(\theta)$ within our finite population framework, since this exceeds the scope of what we can accomplish in the supplement.

We now give more explicit forms for $A(\cdot)$ and $B(\cdot)$ to provide motivation for our specific estimator of $V(\theta_0, \eta_0, \mu_0)$, using the fact that $E\psi_j(\theta_0, \eta_0, \mu_0, 0) = 0$ to simplify.

$$A(\theta_0, \eta_0\mu_0, 0) =$$
$$\lim_{K' \longrightarrow \infty} \begin{pmatrix} -\frac{1}{K'} \sum_{k=1}^{K'} E\left[\nabla_\theta \boldsymbol{\psi}(\mathbf{O}_k, \theta_0)\right] & \begin{matrix} 0 & 0 & 0 \\ \eta_0^{-2} & 0 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & -1 \end{matrix} \\ -\frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \begin{pmatrix} \frac{\partial}{\partial \theta_1} \frac{\text{odds}_{ki}(\theta_0)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} & \cdots & \frac{\partial}{\partial \theta_p} \frac{\text{odds}_{ki}(\theta_0)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} \end{pmatrix} & \\ \mathbf{0}_{1 \times p} & \end{pmatrix}$$

$$B(\theta_0, \eta_0, \mu_0, 0) = \lim_{K' \longrightarrow \infty} \frac{1}{K'} \sum_{k=1}^{K'} E \begin{pmatrix} B_{\theta\theta} & B_{\theta\eta} & B_{\theta\mu} & B_{\theta\widetilde{T}} \\ B_{\theta\eta}^T & B_{\eta\eta} & B_{\eta\mu} & B_{\eta\widetilde{T}} \\ B_{\theta\mu}^T & B_{\eta\mu}^T & B_{\mu\mu} & B_{\mu\widetilde{T}} \\ B_{\theta\widetilde{T}}^T & B_{\eta\widetilde{T}}^T & B_{\mu\widetilde{T}}^T & B_{\widetilde{T}\widetilde{T}} \end{pmatrix}$$

where

$$B_{\theta\theta} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)^T\right\}$$

$$B_{\theta\eta} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\left(\eta_0^{-1} - 1\{n_k>1\}\right)\right\} = -\frac{\eta_0}{K'}\sum_{k=1}^{K} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\right\}$$

$$B_{\theta\mu} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\eta_0 1\{n_k>1\}\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]\right\}$$
$$= \frac{\eta_0}{K'}\sum_{k=1}^{K} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]\right\}$$

$$B_{\theta\widetilde{T}} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\eta_0 1\{n_k>1\}\left[\widetilde{T} - \sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1} + \mu\right]\right\}$$
$$= \frac{\eta_0}{K'}\sum_{k=1}^{K} E\left\{\boldsymbol{\psi}(\mathbf{O}_k,\theta_0)\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1}\right]\right\}$$

$$B_{\eta\eta} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\left(\eta_0^{-1} - 1\{n_k>1\}\right)\right\} = \eta_0^{-2} + \frac{K}{K'}(1-2\eta_0^{-1})$$

$$B_{\eta\mu} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{1\{n_k>1\}(1-\eta_0)\left[\mu - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]\right\}$$
$$= \frac{1-\eta_0}{K'}\sum_{k=1}^{K}\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]$$

$$B_{\eta\widetilde{T}} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{1\{n_k>1\}(1-\eta_0)\left[\widetilde{T} - \sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1} + \mu\right]\right\}$$
$$= \frac{1-\eta_0}{K'}\sum_{k=1}^{K}\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]$$

$$B_{\mu\mu} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\eta_0^2 1\{n_k>1\}\left[\mu - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]^2\right\}$$
$$= \frac{\eta_0^2}{K'}\sum_{k=1}^{K}\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]^2$$

$$B_{\mu\widetilde{T}} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\eta_0^2 1\{n_k>1\}\left[\mu - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]\left[\widetilde{T} - \sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1} + \mu\right]\right\}$$
$$= \frac{\eta_0^2}{K'}\sum_{k=1}^{K}\left[\mu_0 - \sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right)\right]^2$$

$$B_{\widetilde{T}\widetilde{T}} = \frac{1}{K'}\sum_{k=1}^{K'} E\left\{\eta_0^2 1\{n_k>1\}\left[\widetilde{T} - \sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1} + \mu\right]^2\right\}$$
$$= \frac{\eta_0^2}{K'}\sum_{k=1}^{K}\left\{\mu_0^2 - 2\mu_0\sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot \text{odds}_{ki}(\theta_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)} - \frac{1}{n_k-1}\right) + E\left[\sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}-1}{n_k-1}\right]^2\right\}$$

For ease of presentation we let $A(\theta_0)$ and $B(\theta_0)$ be analogous to $A(\theta_0,\eta_0,\mu_0,0)$ and $B(\theta_0,\eta_0,\mu_0,0)$ for the smaller M-estimation problem of fitting the propensity score alone,

and we define $A(\theta_0, \eta_0)$, $B(\theta_0, \eta_0)$, $A(\theta_0, \eta_0, \mu_0)$ and $B(\theta_0, \eta_0, \mu_0)$ analogously. We now invert $A(\theta_0, \eta_0, \mu_0, 0)$. We proceed by representing it as a block matrix:

$$A(\theta_0, \eta_0, \mu_0, 0)^{-1} = \begin{pmatrix} A(\theta_0, \eta_0, \mu_0) & \mathbf{0} \\ (\mathbf{0}, -1) & -1 \end{pmatrix}^{-1} = \begin{pmatrix} A(\theta_0, \eta_0, \mu_0)^{-1} & \mathbf{0} \\ (\mathbf{0}, -1)A(\theta_0, \eta_0, \mu_0)^{-1} & -1 \end{pmatrix}$$

We use a similar strategy to compute $A(\theta_0, \eta_0, \mu_0)^{-1}$:

$$\begin{aligned} A(\theta_0, \eta_0, \mu_0)^{-1} &= \begin{pmatrix} A(\theta_0, \eta_0) & \mathbf{0} \\ \left(\lim_{K' \longrightarrow \infty} \frac{\eta_0}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0), 0\right) & -1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} A(\theta_0, \eta_0)^{-1} & \mathbf{0} \\ \left(\lim_{K' \longrightarrow \infty} \frac{\eta_0}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0), 0\right) A(\theta_0, \eta_0)^{-1} & -1 \end{pmatrix} \end{aligned}$$

where $\mathbf{d}_k(\theta_0) = \left(\frac{\partial}{\partial \theta_1} \frac{\text{odds}_{ki}(\theta_0)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} \quad \cdots \quad \frac{\partial}{\partial \theta_p} \frac{\text{odds}_{ki}(\theta_0)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} \right)$ for each $k = 1, \ldots, K$. Finally, we compute $A(\theta_0, \eta_0)^{-1}$:

$$A(\theta_0, \eta_0)^{-1} = \begin{pmatrix} A(\theta_0) & \mathbf{0} \\ \mathbf{0} & \eta_0^{-2} \end{pmatrix}^{-1} = \begin{pmatrix} A(\theta_0)^{-1} & \mathbf{0} \\ \mathbf{0} & \eta_0^2 \end{pmatrix}$$

We now evaluate our quantity of interest, entry $((p+3), (p+3))$ of the matrix $V(\theta_0, \eta_0, \mu_0)$, which we will denote $v_{\widetilde{T}}(\theta_0, \eta_0, \mu_0)$.

$$v_{\widetilde{T}}(\theta_0, \eta_0, \mu_0) = \mathbf{a}(\theta_0) B(\theta_0, \eta_0, \mu_0, 0) \mathbf{a}(\theta_0)^T$$

where

$$\mathbf{a}(\theta_0) = \left(-\lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) A(\theta_0)^{-1} \quad 0 \quad 1 \quad -1 \right).$$

Expanding this, we obtain the following:

$$
\begin{aligned}
&v_{\widetilde{T}}(\theta_0, \eta_0, \mu_0) = \\
&\left\{ \lim_{K' \longrightarrow \infty} \frac{1}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} B(\theta_0) [A(\theta_0)^{-1}]^T \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\}^T \\
&- 2 \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} \Bigg\{ \\
&\qquad \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} E \left\{ \boldsymbol{\psi}(\mathbf{O}_k, \theta_0) \left[\mu - \sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k \cdot \text{odds}_{ki}(\theta_0)}{(n_k - 1) \sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} - \frac{1}{n_k - 1} \right) \right] \right\} \Bigg\} \\
&- 2 \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} E \left\{ \boldsymbol{\psi}(\mathbf{O}_k, \theta_0) \left[\mu - \sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1} \right] \right\} \right\} \\
&+ \lim_{K' \longrightarrow \infty} \frac{\eta_0^2}{K'} \sum_{k=1}^{K} E \left[\sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1} \right]^2 - \lim_{K' \longrightarrow \infty} \frac{\eta_0^2}{K'} \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k \cdot \text{odds}_{ki}(\theta_0)}{(n_k - 1) \sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} - \frac{1}{n_k - 1} \right) \right]^2 \\
&= \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} B(\theta_0) [A(\theta_0)^{-1}]^T \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\}^T \\
&- 2 \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} \Bigg\{ \\
&\qquad \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} E \left\{ \boldsymbol{\psi}(\mathbf{O}_k, \theta_0) \left[\mu - \sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k \cdot \text{odds}_{ki}(\theta_0)}{(n_k - 1) \sum_{j=1}^{n_k} \text{odds}_{kj}(\theta_0)} - \frac{1}{n_k - 1} \right) \right] \right\} \Bigg\} \\
&- 2 \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k}{n_k - 1} \mathbf{d}_k(\theta_0) \right\} A(\theta_0)^{-1} \left\{ \lim_{K' \longrightarrow \infty} \frac{\eta_0}{K'} \sum_{k=1}^{K} E \left\{ \boldsymbol{\psi}(\mathbf{O}_k, \theta_0) \left[\mu - \sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1} \right] \right\} \right\} \\
&+ \lim_{K' \longrightarrow \infty} \frac{\eta_0^2}{K'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\frac{n_k}{n_k - 1} \right)^2 Y_{ki} p_{ki} \left\{ Y_{ki}(1 - p_{ki}) - \sum_{j \neq i}^{n_k} Y_{kj} p_{kj} \right\}
\end{aligned}
$$

Notice that the last term is identical to the variance of the original difference-in-means estimator T when the true propensity scores are known, derived above (after normalizing both sides by K). While $v_{\widetilde{T}}(\theta_0, \eta_0, \mu_0)$ cannot be calculated in practice without knowledge of the true parameters, the moments of the derivatives of the $\boldsymbol{\psi}$ functions, and the limits in K, we can estimate it by substituting estimated parameters and sample moments/estimates as follows. For convenience, we let $\mathbf{O}_k^{\ell} = \{\mathbf{Y}_k, \mathbf{Z}_k^{\ell}, \mathbf{X}_k\}$ where $\mathbf{Z}_k^{\ell}$ be the length-n_k vector containing a 1 at element ℓ and zeroes otherwise. Note also that to estimate the variance of $\widetilde{T}$, rather than the variance of $\sqrt{K'}\widetilde{T}$, we must also divide our estimate for $v_{\widetilde{T}}(\theta_0, \eta_0, \mu_0)$ by K'.

$$\widehat{Var}(\widetilde{T}) = \frac{1}{K'}\widehat{v}_{\widetilde{T}}(\widehat{\theta},\widehat{\eta},\widehat{\mu}) =$$

$$\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k}{n_k-1}\mathbf{d}_k(\widehat{\theta}_0)\right\}\frac{A_n(\widehat{\theta}_0)^{-1}B_n(\widehat{\theta}_0)[A_n(\widehat{\theta}_0)^{-1}]^T}{K'}\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k}{n_k-1}\mathbf{d}_k(\widehat{\theta}_0)\right\}^T$$

$$-\frac{2}{K'}\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k}{n_k-1}\mathbf{d}_k(\widehat{\theta}_0)\right\}A_n(\widehat{\theta}_0)^{-1}\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{n_k}\frac{\text{odds}_{k\ell}(\widehat{\theta}_0)}{\sum_{j=1}^{n_k}\text{odds}_{kj}(\widehat{\theta}_0)}\boldsymbol{\psi}(\mathbf{O}_k^\ell,\widehat{\theta}_0)\Bigg[\right.$$

$$\left.\frac{1}{K}\sum_{k'=1}^{K}\sum_{i=1}^{n_{k'}} Y_{ki}\left(\frac{n_{k'}\cdot\text{odds}_{k'i}(\theta)}{(n_{k'}-1)\sum_{j=1}^{n_{k'}}\text{odds}_{k'j}(\theta)}-\frac{1}{n_{k'}-1}\right)-\sum_{i=1}^{n_k} Y_{ki}\left(\frac{n_k\cdot\text{odds}_{ki}(\widehat{\theta}_0)}{(n_k-1)\sum_{j=1}^{n_k}\text{odds}_{kj}(\widehat{\theta}_0)}-\frac{1}{n_k-1}\right)\Bigg]\right\}$$

$$-\frac{2}{K'}\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k} Y_{ki}\frac{n_k}{n_k-1}\mathbf{d}_k(\widehat{\theta}_0)\right\}A_n(\widehat{\theta}_0)^{-1}\left\{\frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{n_k}\frac{\text{odds}_{k\ell}(\widehat{\theta}_0)}{\sum_{j=1}^{n_k}\text{odds}_{kj}(\widehat{\theta}_0)}\boldsymbol{\psi}(\mathbf{O}_k^\ell,\widehat{\theta}_0)\Bigg[\right.$$

$$\left.\frac{1}{K}\sum_{k'=1}^{K}\sum_{i=1}^{n_{k'}} Y_{ki}\left(\frac{n_{k'}\cdot\text{odds}_{k'i}(\theta)}{(n_{k'}-1)\sum_{j=1}^{n_{k'}}\text{odds}_{k'j}(\theta)}-\frac{1}{n_{k'}-1}\right)-\sum_{i=1}^{n_k} Y_{ki}\frac{n_k Z_{ki}^\ell-1}{n_k-1}\Bigg]\right\}$$

$$+\frac{1}{K^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\frac{n_k}{n_k-1}\right)^2 Y_{ki}\left(\frac{\text{odds}_{ki}(\widehat{\theta}_0)}{\sum_{\ell=1}^{n_k}\text{odds}_{k\ell}(\widehat{\theta}_0)}\right)\left\{Y_{ki}\left[1-\left(\frac{\text{odds}_{ki}(\widehat{\theta}_0)}{\sum_{\ell=1}^{n_k}\text{odds}_{k\ell}(\widehat{\theta}_0)}\right)\right]-\sum_{j\neq i}^{n_k} Y_{kj}\left(\frac{\text{odds}_{kj}(\widehat{\theta}_0)}{\sum_{\ell=1}^{n_k}\text{odds}_{k\ell}(\widehat{\theta}_0)}\right)\right\}$$

We now give a closed form for the ℓth element of $\mathbf{d}_k(\theta_0)$, i.e. $\frac{\partial}{\partial\theta_\ell}\frac{\text{odds}_{ki}(\theta_0)}{\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta_0)}$, in the special case of the logistic regression model defined by $\log(P(Z=1|X)/P(Z=0|X)) = X\theta$ where the first column of X consists entirely of 1s (and where we index the ℓth column of X_{ki} as $X_{\ell ki}$).

$$\frac{\text{odds}_{ki}(\theta)}{\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta)} = \frac{\exp(X_{ki}\theta)}{\sum_{j=1}^{n_k}\exp(X_{kj}\theta)} = \frac{1}{\sum_{j=1}^{n_k}\exp[(X_{kj}-X_{ki})\theta]}$$

$$\frac{\partial}{\partial\theta_\ell}\frac{\text{odds}_{ki}(\theta)}{\sum_{j=1}^{n_k}\text{odds}_{kj}(\theta)} = -\frac{1}{\left\{\sum_{j=1}^{n_k}\exp[(X_{kj}-X_{ki})\theta]\right\}^2}\sum_{j=1}^{n_k}\exp[(X_{kj}-X_{ki})\theta](X_{\ell kj}-X_{\ell ki})$$

This quantity will be zero for $\ell = 1$ since $X_{1kj} = X_{1ki} = 1$ for all i, j. The estimating equations $\boldsymbol{\psi}(\mathbf{O}_k, \widehat{\theta}_0)$ are discussed in many standard textbooks and have the following form in our notation:

$$\psi_j(\mathbf{O}_k,\widehat{\theta}_0) = \frac{1}{K'}\sum_{k=1}^{K'}\sum_{i=1}^{n_k}\left(Z_{ki}-\frac{1}{1+\exp(-X_{ki}\theta)}\right)X_{j,ki}$$

where $X_{1,ki} = 1$ for all k, i. In turn, the gradient of ψ_j has the following form:

$$\nabla\psi_j(\mathbf{O}_k,\widehat{\theta}_0) = -\sum_{i=1}^{n_k}\left(\ \frac{\exp(-X_{ki}\theta)X_{1,ki}X_{j,ki}}{[1+\exp(-X_{ki}\theta)]^2},\ \ \cdots\ \ ,\frac{\exp(-X_{ki}\theta)X_{j,ki}^2}{[1+\exp(-X_{ki}\theta)]^2},\ \ \cdots\ \ ,\frac{\exp(-X_{ki}\theta)X_{p,ki}X_{j,ki}}{[1+\exp(-X_{ki}\theta)]^2}\ \right)$$

Sample averages of these gradients (multiplied by -1 and evaluated at estimated values of θ_0) form the rows of the matrix $A_n(\widehat{\theta}_0)$ in the variance estimate given above. Note also

that the term $A_n(\widehat{\theta}_0)^{-1}B_n(\widehat{\theta}_0)[A_n(\widehat{\theta}_0)^{-1}]^T$ is simply an estimate of the variance-covariance matrix for propensity score estimation; in our framework, in which finite populations exhibit internal matched-set structure and in which individual O_k terms may correspond to clusters of units rather than distinct subjects, this is a clustered covariance estimate such as that of Liang and Zeger [53] with clustering over matched sets (and unmatched individuals treated as their own clusters).

The above all assumes that the propensity score is estimated in the same sample in which the match is conducted. Note that Theorems 1-2 instead assume the propensity score estimation is done in a different sample. In this case the variance estimation problem becomes simpler and does not actually require an M-estimation framework, since the propensity score estimates and the Z_{ki}s are independent. Instead, the delta method can be used to calculate the variance of $\widehat{\mu}$ and this variance can be added to the estimated variance of T to obtain an overall variance. More specifically, we can write:

$$h(\theta) = \sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k \cdot \text{odds}_{ki}(\theta)}{(n_k - 1)\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta)} - \frac{1}{n_k - 1} \right)$$

Let Σ be the asymptotic variance of $\widehat{\theta}$. Then by the delta method,

$$\text{Var}(h(\theta)) \approx \nabla h(\theta)^T \Sigma \nabla h(\theta).$$

where

$$\nabla h(\theta) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \left(\frac{n_k}{n_k - 1} \right) \left(\frac{\partial}{\partial \theta_1} \frac{\text{odds}_{ki}(\theta)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta)} \quad \cdots \quad \frac{\partial}{\partial \theta_p} \frac{\text{odds}_{ki}(\theta)}{\sum_{j=1}^{n_k} \text{odds}_{kj}(\theta)} \right)$$

This quantity may be estimated by substituting $\widehat{\theta}$ for θ. Note that the resulting variance estimate is almost identical to $\widehat{\text{Var}}(\widetilde{T})$ as given above for the case of in-sample estimation, except with no cross-terms and with matrix Σ, which must be estimated in the pilot sample, in place of the in-sample covariance matrix for $\widehat{\theta}$.

C.5 Additional simulation results

We provide further detail on the simulation settings explored in Section 6.2 of the main manuscript and give results for several alternate settings. First, for the primary setting considered in the main manuscript, Figure C.1 shows smoothed density plots comparing the true propensity score distribution across treatment and control groups for sample draws from each of six different simulation settings affecting the treatment model. The density plots are scaled by the relative size of the two groups; in large samples, regions of the plots where the control density exceeds the treated density indicate propensity score values at which near-exact matching on the propensity score is possible, while regions with larger treated density suggest regions in which either poorer matches must be accepted or treated units must be

trimmed [51]; the number given in the title of the plot is an estimate of the proportion of the probability mass in the treated group that overlaps with the scaled control density, indicating easy matchability. In all the settings shown here, most treated units can be matched closely but in most cases there is some region in which matching is more difficult, ensuring that inexact matches occur systematically. Of course these plots are smoothed depictions that mask the underlying discreteness of the data, particularly in the $n = 100$ case, but they suggest the general difficulty of the matching problem.

The next set of figures repeat the analyses given in Figure 1 from the main manuscript for several other simulation settings as discussed in Section 6.2 . In particular, Figures C.2 - C.3 give results for a weak propensity score setting where the treated and control groups are more similar than in the primary simulations; Figure C.4 gives results averaging only over simulations in which good covariate balance was achieved in the sense that absolute standardized differences in means for all measured covariates were under 0.2 after matching; and Figure C.5 gives results for a setting in which propensity scores were estimated in a large external sample. The weak-signal results show similar patterns to the primary simulations but the size of type I error violations is substantially reduced and uniform randomization inference sometimes achieves Type I error control. The density plots reveal that the weak-signal settings correspond to a setting in which the distribution of true propensity scores is very similar between groups; when propensity score distributions show such similarity in real datasets this suggests that uniform randomization inference may be warranted. The other two simulation settings produce patterns of Type I error rates almost identical to those in Figure 1 in the main manuscript.

Finally, we examine the quality of point estimates constructed using the maximum p-value procedure of Section 3.3 in the main manuscript. We repeat the simulations described in Section 6.2 with a strong propensity score signal and without introducing regression adjustment, calipers, uniform inference, or nonlinearity in the outcome model; we also add a treatment effect of magnitude 1 and compute both the maximum p-value point estimate and the traditional difference-in-means. Table C.2 reports results. The maximum p-value estimates are uniformly preferable to the difference-in-means estimates, and except in a few cases with Z-dependence they are very close to the true effect.

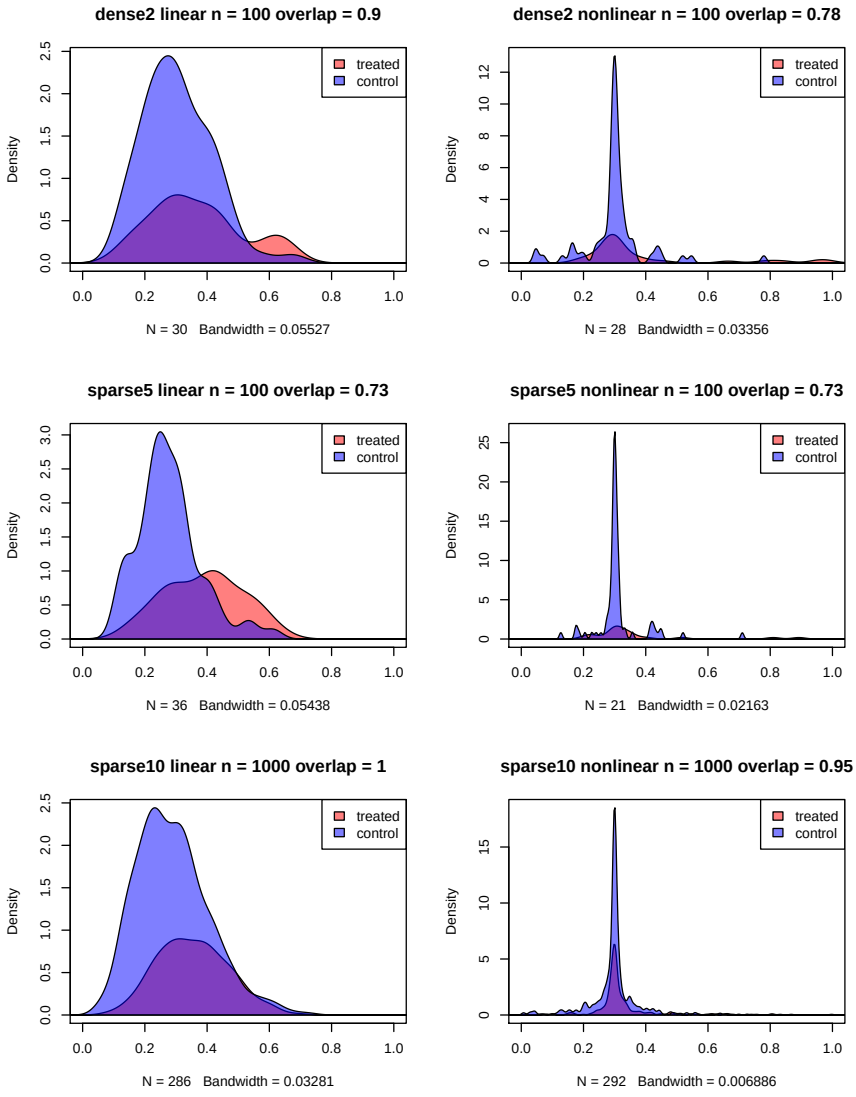

Figure C.1: Smoothed and scaled density plots of one random draw each from six different simulation settings for the treatment model. The first and second columns contrast weak and strong propensity score signals ($\tau = 0.2$ vs. $\tau = 0.6$), and the three rows contrast three different sizes for the initial dataset: $(n, p) = (100, 2), (100, 5)$, and (1000,10) respectively.

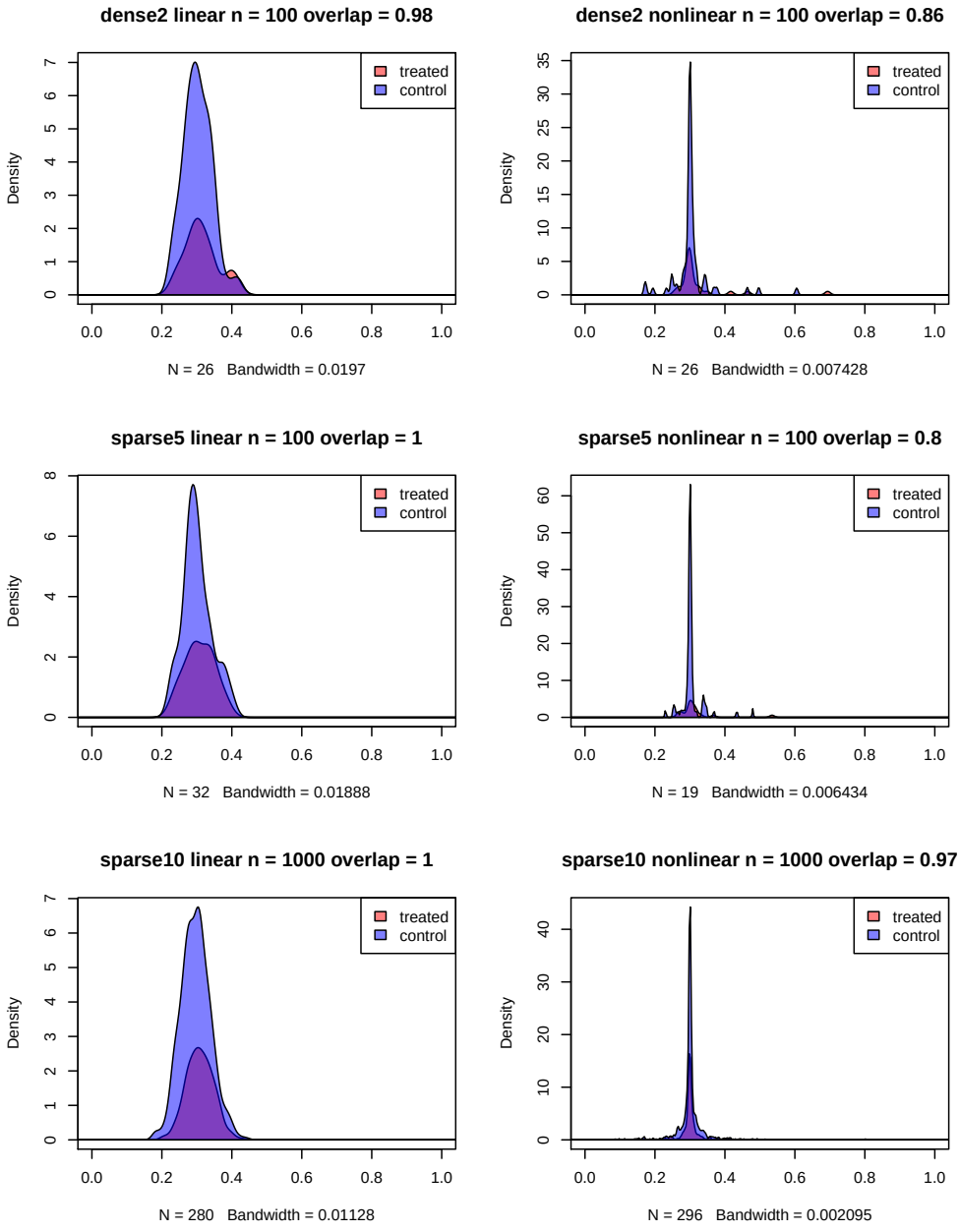

Figure C.2: Relative density of treated and control groups by propensity score value under a weak propensity score signal. For more detail, see caption to Figure C.1.

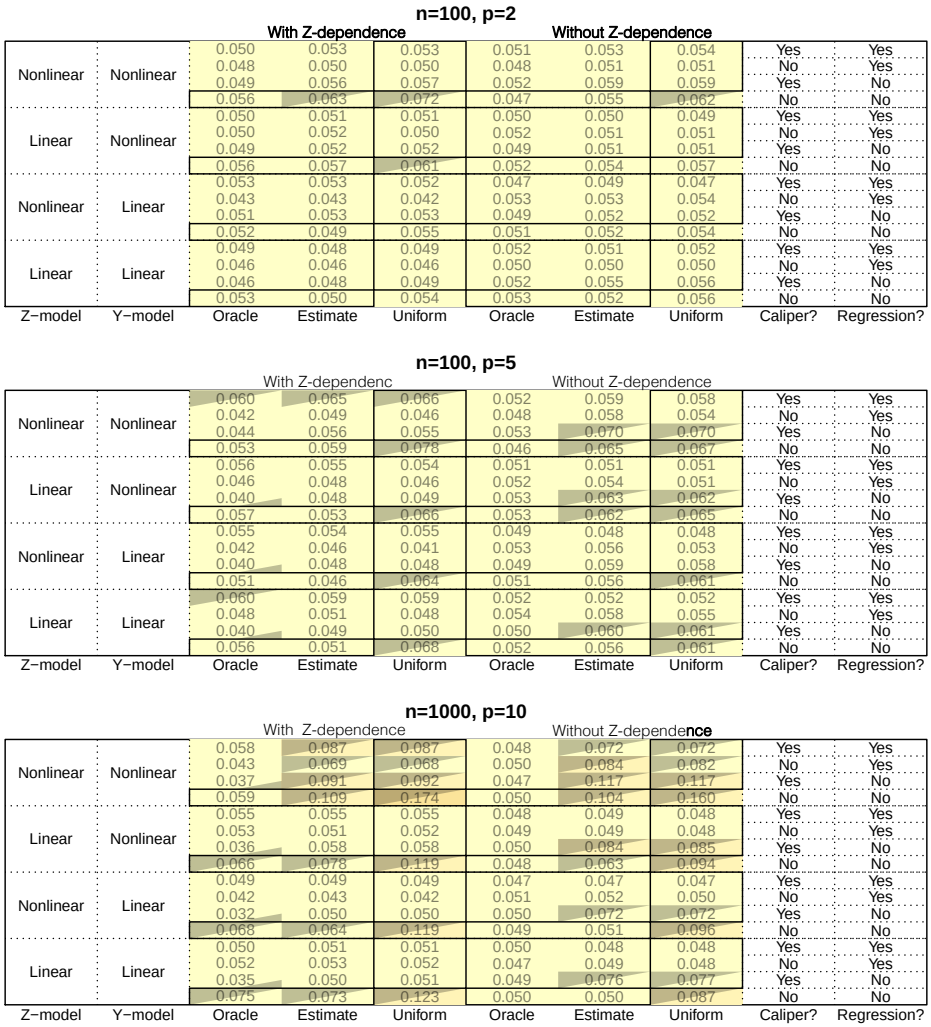

Figure C.3: Type I error results under weak propensity score signal for uniform and covariate-adaptive inference across multiple simulation settings. Each of the three tables corresponds to a separate dataset size. Within each table the first three columns contrast three inferential approaches when the subjects are subject to Z-dependence; the last three columns do the same comparison when treatment assignments within matched sets are permuted after matching to eliminate Z-dependence. The rows of the table demonstrate different combinations of calipers and regression adjustment and correct or incorrect specification of treatment and outcome models. Numbers give type I error rates with colors associated to their magnitude; triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05 (under a Bonferroni correction scaled to the number of results across the entire figure), with a large upper triangle indicating a positive z-statistic and a small lower triangle indicating a negative z-statistic.

n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.055	0.058	0.057	0.055	0.056	0.056	Yes	Yes
		0.047	0.051	0.050	0.056	0.059	0.059	No	Yes
		0.051	0.057	0.057	0.052	0.060	0.061	Yes	No
		0.061	0.078	0.094	0.054	0.072	0.087	No	No
Linear	Nonlinear	0.048	0.048	0.049	0.052	0.051	0.051	Yes	Yes
		0.043	0.042	0.040	0.054	0.052	0.051	No	Yes
		0.044	0.048	0.049	0.053	0.056	0.056	Yes	No
		0.051	0.057	0.067	0.054	0.058	0.064	No	No
Nonlinear	Linear	0.052	0.051	0.052	0.050	0.050	0.049	Yes	Yes
		0.034	0.035	0.033	0.046	0.046	0.045	No	Yes
		0.051	0.052	0.052	0.049	0.051	0.050	Yes	No
		0.053	0.053	0.070	0.049	0.050	0.062	No	No
Linear	Linear	0.050	0.049	0.050	0.053	0.053	0.052	Yes	Yes
		0.048	0.046	0.047	0.051	0.053	0.052	No	Yes
		0.049	0.050	0.052	0.053	0.056	0.056	Yes	No
		0.061	0.060	0.072	0.054	0.053	0.063	No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.068	0.080	0.079	0.048	0.054	0.055	Yes	Yes
		0.059	0.071	0.067	0.057	0.068	0.064	No	Yes
		0.044	0.068	0.071	0.050	0.080	0.080	Yes	No
		0.048	0.067	0.087	0.051	0.086	0.106	No	No
Linear	Nonlinear	0.061	0.059	0.061	0.051	0.051	0.050	Yes	Yes
		0.057	0.054	0.051	0.045	0.048	0.047	No	Yes
		0.043	0.054	0.055	0.048	0.070	0.071	Yes	No
		0.047	0.046	0.061	0.043	0.055	0.064	No	No
Nonlinear	Linear	0.061	0.056	0.057	0.050	0.051	0.049	Yes	Yes
		0.048	0.046	0.045	0.052	0.053	0.051	No	Yes
		0.041	0.049	0.050	0.049	0.063	0.063	Yes	No
		0.049	0.049	0.066	0.051	0.059	0.088	No	No
Linear	Linear	0.063	0.061	0.059	0.051	0.050	0.048	Yes	Yes
		0.064	0.065	0.060	0.034	0.036	0.034	No	Yes
		0.041	0.046	0.050	0.049	0.072	0.074	Yes	No
		0.051	0.050	0.074	0.046	0.058	0.072	No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.052	0.080	0.080	0.048	0.074	0.074	Yes	Yes
		0.037	0.084	0.080	0.047	0.115	0.107	No	Yes
		0.041	0.098	0.100	0.052	0.116	0.118	Yes	No
		0.053	0.199	0.422	0.047	0.190	0.410	No	No
Linear	Nonlinear	0.048	0.049	0.050	0.051	0.051	0.051	Yes	Yes
		0.045	0.046	0.045	0.061	0.059	0.056	No	Yes
		0.037	0.059	0.063	0.051	0.081	0.084	Yes	No
		0.065	0.106	0.245	0.060	0.095	0.215	No	No
Nonlinear	Linear	0.044	0.044	0.043	0.051	0.050	0.050	Yes	Yes
		0.027	0.027	0.024	0.054	0.054	0.051	No	Yes
		0.042	0.054	0.059	0.051	0.067	0.070	Yes	No
		0.053	0.069	0.211	0.062	0.068	0.230	No	No
Linear	Linear	0.050	0.050	0.050	0.052	0.051	0.051	Yes	Yes
		0.050	0.049	0.044	0.059	0.056	0.052	No	Yes
		0.048	0.061	0.065	0.050	0.067	0.070	Yes	No
		0.082	0.099	0.250	0.061	0.070	0.210	No	No

Figure C.4: Type I error results conditional on good covariate balance for uniform and covariate-adaptive inference across multiple simulation settings. For a more thorough description of how the tables are organized see the caption to Figure C.3.

n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.050	0.048	0.049	0.047	0.047	0.047	Yes	Yes
		0.038	0.040	0.040	0.051	0.055	0.054	No	Yes
		0.050	0.051	0.052	0.048	0.050	0.051	Yes	No
		0.056	0.078	0.104	0.053	0.071	0.089	No	No
Linear	Nonlinear	0.044	0.044	0.044	0.051	0.050	0.050	Yes	Yes
		0.044	0.042	0.041	0.052	0.051	0.049	No	Yes
		0.049	0.048	0.048	0.051	0.050	0.051	Yes	No
		0.059	0.067	0.080	0.051	0.057	0.065	No	No
Nonlinear	Linear	0.047	0.047	0.048	0.054	0.053	0.053	Yes	Yes
		0.034	0.034	0.033	0.049	0.050	0.048	No	Yes
		0.049	0.048	0.051	0.056	0.054	0.055	Yes	No
		0.061	0.062	0.088	0.050	0.050	0.064	No	No
Linear	Linear	0.047	0.047	0.047	0.051	0.050	0.051	Yes	Yes
		0.041	0.041	0.041	0.051	0.051	0.049	No	Yes
		0.052	0.051	0.052	0.052	0.052	0.053	Yes	No
		0.069	0.068	0.090	0.051	0.051	0.060	No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.046	0.047	0.047	0.049	0.048	0.048	Yes	Yes
		0.040	0.049	0.048	0.053	0.063	0.060	No	Yes
		0.049	0.050	0.050	0.050	0.050	0.051	Yes	No
		0.060	0.086	0.130	0.052	0.073	0.112	No	No
Linear	Nonlinear	0.047	0.048	0.047	0.058	0.057	0.057	Yes	Yes
		0.042	0.039	0.037	0.047	0.046	0.045	No	Yes
		0.050	0.051	0.051	0.052	0.053	0.054	Yes	No
		0.060	0.068	0.104	0.047	0.055	0.078	No	No
Nonlinear	Linear	0.045	0.045	0.046	0.048	0.047	0.048	Yes	Yes
		0.034	0.034	0.031	0.051	0.051	0.047	No	Yes
		0.050	0.049	0.051	0.049	0.049	0.050	Yes	No
		0.063	0.062	0.114	0.051	0.052	0.091	No	No
Linear	Linear	0.045	0.046	0.046	0.054	0.053	0.054	Yes	Yes
		0.041	0.041	0.038	0.044	0.045	0.041	No	Yes
		0.047	0.047	0.049	0.053	0.054	0.054	Yes	No
		0.076	0.076	0.127	0.046	0.046	0.080	No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.052	0.061	0.060	0.046	0.051	0.050	Yes	Yes
		0.033	0.082	0.077	0.049	0.115	0.108	No	Yes
		0.063	0.074	0.075	0.048	0.058	0.059	Yes	No
		0.079	0.262	0.543	0.050	0.191	0.453	No	No
Linear	Nonlinear	0.046	0.048	0.048	0.047	0.046	0.045	Yes	Yes
		0.047	0.043	0.042	0.050	0.048	0.047	No	Yes
		0.055	0.059	0.062	0.048	0.050	0.051	Yes	No
		0.095	0.140	0.333	0.052	0.087	0.224	No	No
Nonlinear	Linear	0.048	0.047	0.048	0.049	0.051	0.049	Yes	Yes
		0.023	0.023	0.019	0.048	0.049	0.041	No	Yes
		0.061	0.064	0.067	0.051	0.053	0.056	Yes	No
		0.102	0.100	0.402	0.051	0.050	0.265	No	No
Linear	Linear	0.045	0.046	0.047	0.053	0.052	0.052	Yes	Yes
		0.043	0.042	0.040	0.050	0.050	0.046	No	Yes
		0.062	0.064	0.070	0.052	0.054	0.056	Yes	No
		0.144	0.146	0.435	0.047	0.047	0.224	No	No

Figure C.5: Type I error results for uniform and covariate-adaptive inference across multiple simulation settings, using out-of-sample propensity scores. For a more thorough description of how the tables are organized see the caption to Figure C.3.

Z-model	Y-model	Z-dependence?	τ	Maxp.est	IPW.est	DiM.est
Linear	Linear	No	0	-0.076(0.906)	-0.110(1.119)	0.468(0.970)
Linear	Linear	Yes	0	0.825(1.070)	0.842(1.044)	1.284(1.255)
Linear	Nonlinear	No	0	-0.250(1.620)	-0.221(2.548)	0.792(1.693)
Linear	Nonlinear	Yes	0	0.598(1.530)	0.829(1.741)	1.582(1.859)
Nonlinear	Linear	No	0	0.328(0.852)	0.364(0.878)	0.679(0.898)
Nonlinear	Linear	Yes	0	0.806(0.883)	0.841(0.881)	1.124(1.038)
Nonlinear	Nonlinear	No	0	0.882(1.344)	1.125(1.489)	1.696(1.743)
Nonlinear	Nonlinear	Yes	0	1.357(1.380)	1.602(1.507)	2.134(1.870)
Linear	Linear	No	1	0.924 (0.906)	0.899 (1.086)	1.468 (0.907)
Linear	Linear	Yes	1	1.825 (1.07)	1.806 (1.01)	2.284 (1.255)
Linear	Nonlinear	No	1	0.75 (1.62)	0.788 (2.518)	1.792 (1.693)
Linear	Nonlinear	Yes	1	1.598 (1.53)	1.793 (1.718)	2.582 (1.859)
Nonlinear	Linear	No	1	1.328 (0.852)	1.36 (0.865)	1.679 (0.898)
Nonlinear	Linear	Yes	1	1.806 (0.883)	1.817 (0.864)	2.124 (1.038)
Nonlinear	Nonlinear	No	1	1.882 (1.344)	2.121 (1.476)	2.696 (1.743)
Nonlinear	Nonlinear	Yes	1	2.357 (1.38)	2.579 (1.494)	3.134 (1.87)
Linear	Linear	No	10	9.924 (0.906)	9.981 (0.819)	10.468 (0.907)
Linear	Linear	Yes	10	10.825 (1.07)	10.482 (0.754)	11.284 (1.255)
Linear	Nonlinear	No	10	9.75 (1.62)	9.87 (2.263)	10.792 (1.693)
Linear	Nonlinear	Yes	10	10.598 (1.53)	10.47 (1.553)	11.582 (1.859)
Nonlinear	Linear	No	10	10.328 (0.852)	10.322 (0.762)	10.679 (0.898)
Nonlinear	Linear	Yes	10	10.806 (0.883)	10.603 (0.734)	11.124 (1.038)
Nonlinear	Nonlinear	No	10	10.882 (1.344)	11.084 (1.374)	11.696 (1.743)
Nonlinear	Nonlinear	Yes	10	11.357 (1.38)	11.365 (1.389)	12.134 (1.87)
Linear	Linear	No	100	99.924 (0.906)	100.798 (3.224)	100.468 (0.907)
Linear	Linear	Yes	100	100.825 (1.07)	97.246 (3.535)	101.284 (1.255)
Linear	Nonlinear	No	100	99.75 (1.62)	100.687 (2.786)	100.792 (1.693)
Linear	Nonlinear	Yes	100	100.598 (1.53)	97.233 (3.668)	101.582 (1.859)
Nonlinear	Linear	No	100	100.328 (0.852)	99.946 (1.805)	100.679 (0.898)
Nonlinear	Linear	Yes	100	100.806 (0.883)	98.467 (2.15)	101.124 (1.038)
Nonlinear	Nonlinear	No	100	100.882 (1.344)	100.708 (1.822)	101.696 (1.743)
Nonlinear	Nonlinear	Yes	100	101.357 (1.38)	99.229 (2.249)	102.134 (1.87)

Table C.1: Point estimates from three estimation approaches under a true constant additive effect model: the maximum p-value strategy, inverse propensity weighting, and naïve difference-in-means. Numbers are averaged over 1000 simulation replicates with $n = 100$, $p = 2$, strong propensity signal, and propensity scores estimated in-sample (as in Section 6.3 of the main manuscript). Standard deviations across the 1000 replicates are shown in brackets. The column τ gives the true value of the constant additive effect.

					Point Estimates	
n	p	Z-model	Z-dependence?	Propensity Score	Max P-value	Diff. in Means
100	2	Linear	Yes	Estimated	1.09	1.14
100	2	Linear	Yes	Oracle	1.09	1.14
100	2	Linear	No	Estimated	0.99	1.04
100	2	Linear	No	Oracle	1.00	1.04
100	2	Nonlinear	Yes	Estimated	1.09	1.12
100	2	Nonlinear	Yes	Oracle	1.06	1.12
100	2	Nonlinear	No	Estimated	1.03	1.07
100	2	Nonlinear	No	Oracle	1.00	1.07
100	5	Linear	Yes	Estimated	1.11	1.25
100	5	Linear	Yes	Oracle	1.12	1.25
100	5	Linear	No	Estimated	0.97	1.13
100	5	Linear	No	Oracle	0.99	1.13
100	5	Nonlinear	Yes	Estimated	1.11	1.21
100	5	Nonlinear	Yes	Oracle	1.07	1.21
100	5	Nonlinear	No	Estimated	1.03	1.14
100	5	Nonlinear	No	Oracle	1.01	1.14
1000	10	Linear	Yes	Estimated	1.10	1.25
1000	10	Linear	Yes	Oracle	1.11	1.25
1000	10	Linear	No	Estimated	1.00	1.15
1000	10	Linear	No	Oracle	1.00	1.15
1000	10	Nonlinear	Yes	Estimated	1.10	1.20
1000	10	Nonlinear	Yes	Oracle	1.06	1.20
1000	10	Nonlinear	No	Estimated	1.04	1.15
1000	10	Nonlinear	No	Oracle	1.00	1.15

Table C.2: Point estimates from the maximum p-value strategy described in Section 3.3 in the main manuscript vs. difference-in-means estimates. The true value of the parameter is 1.