Cambridge Medium (2020), **37**, 1–22

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE TYPE

The Power of Prognosis: Improving Covariate Balance Tests with Outcome Information

Abstract

Scholars often use covariate balance tests to test the validity of natural experiments and related designs. Unfortunately, when covariates are unrelated to potential outcomes, balance on measured covariates tells us nothing about whether key identification conditions are met. We underscore the importance of basing tests on prognostic variables—those most predictive of potential outcomes—and propose a bootstrap balance test that prioritizes informative covariates. We adapt our procedure to regression-discontinuity designs, introduce an open-source software package, and apply the tests to a sample of applied studies. A debate about the randomness of close elections illustrates the importance of incorporating information about covariate prognosis in balance tests.

Keywords: covariate balance tests, placebo outcomes, natural experiments, as-if random, regression discontinuity designs, continuity of potential outcomes, weighting methods

1. Introduction

Methodologists urge researchers to test observable implications of assumptions that facilitate causal inference. The majority of studies using natural experiments and regression-discontinuity designs in three top political science journals present some form of covariate balance test. The logic appears straightforward: if a coin flip determined treatment assignment, as assumed or stipulated by design, pre-treatment covariates or "placebo outcomes" would have the same distribution, in expectation, in the treatment and control groups (Eggers, Tuñón, and Dafoe 2023, Caughey, Dafoe, and Seawright 2017). A statistically insignificant association between treatment and covariates is taken to be consistent with random assignment—an important advantage, if true, for making inferences about causation—while a significant association may suggest a flaw in the design.

Unfortunately, these widely used tests may shed no light on key identification conditions for causal inference. In natural experiments, the key condition is the independence of treatment assignment and potential outcomes.[1] This is sometimes called "as-if" random (Freedman 2009, Dunning 2012). In many regression discontinuity (RD) designs, identification instead depends on the continuity of potential outcome regression functions at a threshold determining treatment assignment. In either case, tests based on pre-treatment covariates can be uninformative.

For example, in a voter mobilization campaign, citizens who are already more likely to vote may decide to talk to canvassers—contrary to the postulate that assignment is as-if random. Yet, men may be as likely to do so as women. This leads to expected balance on gender across the treatment and control groups, but imbalance on prior turnout. If we only have data on gender, we may fail to reject as-if random. Yet, potential responsiveness to the intervention is likely imbalanced across the groups.

Conversely, men might tend to select into treatment, yet selection is unrelated to responsiveness to the intervention. We may then reject as-if random based on a gender imbalance—even if potential outcomes themselves are balanced. Post-treatment turnout may be strongly related to pre-treatment

1. Potential outcomes are the outcomes that would be realized under counterfactual assignment to different treatments (Neyman, Dabrowska, and Speed 1923, Rubin 1974).

turnout, while gender may be unrelated to treatment responsiveness. The key point is that different covariates vary in their informativeness about potential outcomes.

Standard balance tests do not take this variation in covariate informativeness into account. Moreover, in some balance tests, *none* of the measured covariates are prognostic—that is, associated with potential outcomes. Although potential outcomes are partially unobservable—once a unit has been assigned to the control group, we cannot see its potential outcome under treatment (Holland 1986)—a natural experiment makes it possible to assess how well covariates predict potential outcomes, as we describe later. However, it is rare for researchers to provide diagnostics of prognosis in their covariate balance tests. As we demonstrate, the prognosis of covariates used in balance tests in prominent natural experiments and RD designs in political science varies and is often quite low.

We show that tests using irrelevant, non-prognostic covariates lead researchers to over-reject as-if random when it is true or to fail to reject when it is false. Conversely, when measured covariates are minimally sufficient—that is, fully informative about potential outcomes—covariates are imbalanced if and only if potential outcomes are imbalanced. Sufficiency of measured covaraites may be empirically rare. However, we show that even if covariates are not sufficient, researchers can increase the power and specificity of their tests by prioritizing the most informative covariates, among those they measure. Moreover, the tests' performance improves as the joint prognosis of covariates increases.

We therefore argue that the best approach to testing as-if random or continuity of potential outcomes requires two steps. First, analysts should seek to gather data on the most prognostic set of covariates possible. Importantly, they should report diagnostics of covariate prognosis, which researchers now virtually never do. We recommend the R^2 from a regression of measured post-treatment outcomes—e.g., observed potential outcomes under control—on covariates, which we call the "prognosis R^2." Such goodness-of-fit measures are essential because empirically, covariate prognosis varies widely across applications. While pre-treatment (lagged) measures of outcome variables tend to be related to potential outcomes (Imbens and Rubin 2015: 483-4), lagged outcomes may or may not be available; and as we show, in some applications they are not in fact prognostic.

Second, analysts should then prioritize in their tests the particular covariates most associated with potential outcomes. We propose one straightforward approach that upweights the most prognostic individual covariates while downweighting non-prognostic variables. Our prognosis-weighted (alternately, "informativeness-weighted") test of as-if random is based on the treatment-control group difference in the average $\widehat{Y_i(0)}$, the covariate-adjusted potential outcome under control. The test statistic is equivalent to a weighted sum of standardized mean differences for individual covariates, where the weights are measures of prognosis. This approach has the advantage of a close connection to current practice—as it is based on a combination of the covariate-specific test statistics used in standard tests—and, as we show, the weights are readily interpretable as the relative informativeness of different covariates. We also develop a test appropriate for discontinuity designs, in which differences of means are replaced by fitted intercepts to test for continuity of average potential outcomes at the RD threshold (rather than as-if random) (De la Cuesta and Imai 2016, Sekhon and Titiunik 2017).

Our theoretical results, as well as simulations reported in the online Appendix, suggest that these prognosis-weighted tests avoid the problem of irrelevant covariates. By downweighting uninformative variables and upweighting informative ones, the approach simultaneously limits both false positives and false negatives. We also discuss how to adapt our approach for equivalence tests (Hartman and Hidalgo 2018) and develop a design-based approach to hypothesis testing using bootstrapping. We implement all procedures in our forthcoming R package `pwtest`.[2]

After developing the theory, we apply our approach to a sample of published natural experiments and regression-discontinuity designs. We show that both the extent of imbalance and, especially, the degree of covariate prognosis vary across prominent studies. We report p-values for these studies and show graphically how our prognosis-weighted tests project out irrelevant covariates. Finally,

2. Package `pwtest` can be found on https://github.com/[ANONYMIZED]/pwtest. See Appendix Section 6.

we develop a case study of the randomness of close elections, an important topic of recent debate (Caughey and Sekhon 2011, Eggers et al. 2015, De la Cuesta and Imai 2016, Hartman 2021). This controversy underscores the importance of diagnosing prognosis and the usefulness of the specific procedure we propose, which allows researchers to base their conclusions on variables most related to potential outcomes.

Overall, we show that incorporating information about covariate prognosis into balance tests improves on current practice—which ignores covariate informativeness altogether—and can lead to more credible conclusions about whether identifying conditions are met. By basing tests on a single summary test statistic that is a combination individual differences of means or fitted intercepts—i.e., the inputs for standard tests—the tests we propose also provide a new way to address multiple testing concerns (De la Cuesta and Imai 2016). We therefore add to valuable articles on omnibus covariate balance tests (Hansen and Bowers 2008; Caughey, Dafoe, and Seawright 2017; Gagnon-Bartsch and Shem-Tov 2019), which do not however consider covariate progrnosis.[3]

Our proposed approach is one way to increase the informativeness of balance tests and improve current practice. We view it as complementary to other kinds of evidence on identifying conditions. For example, qualitative information often bolsters or undercuts the plausibility of as-if random or of continuity of potential outcomes in natural experiments and RD designs (Dunning 2012). As-if random and continuity—if true—facilitate causal inference using relatively simple, transparent methods (Freedman 2009). Our focus here is therefore on improving tests to assess whether these assumptions hold, rather than on strategies for limiting bias when they fail. Our aim is thus distinct from a literature on optimizing observed balance to estimate treatment effects, through matching or other techniques. For example, Hansen (2008) proposes a "prognostic score" related to our informativeness weights to estimate average treatment effects (see also Stuart, Lee, and Leacy 2013, Leacy and Stuart 2014, and Wainstein 2022). Because our aim is to test as-if random or continuity of potential outcomes—not to limit bias when those conditions fail—we do not propose matching on a propensity score (Leacy and Stuart 2014,Rubin and Thomas 2000), since the independence of treatment assignement propensity and potential outcomes is exactly what we want to test. Our theory and simulations suggest that the performance of tests suffers in exactly those settings where optimizing observed balance can be illusory, i.e., when there are unmeasured covariates associated with treatment and with potential outcomes. In this case, however, our diagnostics—such as the R^2 from our prognosis regression—may suggest the limited power of the tests.

In the next section, we discuss formally why prognosis matters for balance tests. We then develop statistical theory behind our informativeness-weighted approach, and we develop a bootstrap hypothesis test. We also refer to simulation evidence on the test's power and specificity. We then turn to prognosis weighting in practice, applying our approach to a sample of natural experiments and regression discontinuity designs and studying in detail the case of close elections. We conclude by discussing avenues for future research.

2. Are potential outcomes balanced? Why prognosis matters

There are at least two reasons that prognosis of covariates matters for testing as-if random—and also thus why covariates with differing degrees of prognosis should not be "treated equal."

First, the most direct test of this critical condition for causal inference would assess balance of *potential outcomes* across the treatment and control groups (Imbens and Rubin 2015). A direct test of this assumption is impossible—due to the fact that once treatment has occurred, we do not observe potential outcomes under control in a treatment group or potential outcomes under treatment in a

3. Related research in statistics and epidemiology recommends upweighting tests for "important" hypotheses—those most plausibly false—in p-value combinations; see e.g. Fisher (1935), Holm (1979), Benjamini and Hochberg (1997), Kost and McDermott (2002), Westfall (2014), and Genovese, Roeder, and Wasserman (2006). Our approach gives specific content to which hypotheses are most likely to be false in balance tests by upweighting covariates related to potential outcomes.

control group. Yet, a covariate strongly associated with potential outcomes may give us substantial information about this realized balance.

Second and relatedly, if subjects have the opportunity to select into treatment, as in many observational studies, they may do so in a way that reflects the outcomes they would experience under treatment or control (Angrist and Pischke 2009). Prioritizing covariates strongly related to potential outcomes is most likely to detect such violations of the key condition.

2.1 Defining as-if random

We develop our argument formally using a design-based, finite population set-up. Consider a study with a completely enumerated finite population of N units indexed by $i = 1, \ldots, N$ and one treatment and one control condition. Let $Y_i(1)$ and $Y_i(0)$ be potential outcomes—that is, the outcomes for unit i that would be realized under assignment to treatment or control groups, respectively. The causal effect for each unit is $\tau_i = Y_i(1) - Y_i(0)$, while the Average Treatment Effect (ATE) is $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, where the expectation is taken over the draw of a single unit at random from the finite population.[4] The random variable $Z_i \in \{0, 1\}$ denotes treatment assignment, with 0 for the control group and 1 for the treatment group; an $N \times 1$ random vector Z collects the Z_i. The sizes of the treatment and control groups are fixed at n_1 and n_0, respectively, with $n_1 + n_0 = N$.

In natural experiments, the following condition must hold by definition (Dunning 2012):

Assumption 1 *(As-if Random Assignment)* $Z \perp\!\!\!\perp \{Y(1), Y(0)\}$

where $\perp\!\!\!\perp$ denotes "is independent of."[5] In words, treatment is assigned independently of potential outcomes.[6] This ensures, for example, that sicker patients do not go systematically to the treatment group in a drug trial studying health outcomes, or that those less prone to vote do not disproportionately receive a vote-mobilizing intervention. If as-if random holds, the true ATE is estimable using simple, transparent methods (Freedman 1999).

Assumption 1 cannot be directly verified due to the "fundamental problem of causal inference": $\{Y_i(1), Y_i(0)\}$ is not completely observed for any unit (Holland 1986). In a randomized experiment, independence of potential outcomes and treatment assignment is an implication of a chance protocol, often under the control of a researcher (Fisher 1935). In natural experiments, as-if random is held to be an implication of a concrete process that produces a haphazard allocation to treatments, in particular, one that does not depend on units' potential outcomes. This assumption can be the "Achilles Heel" of natural experiments (Dunning 2008). With additional assumptions discussed below, however, it is possible to make as-if random falsifiable.

2.2 Standard balance tests, and two counterexamples

Consider now a set of possible covariates $\mathcal{X}$. We suppose that $\mathcal{X} = \{\mathcal{X}^S, \mathcal{X}^N\}$ ('signal' and 'noise', respectively – compare Liu and Ruan 2020), where $\mathcal{X}^S$ contains information about potential outcomes and $\mathcal{X}^N$ does not. Treat $\mathcal{X}^S$ and $\mathcal{X}^N$ as finite but potentially unobserved. With a slight abuse of notation,[7]

$$\mathcal{X}^S \not\perp\!\!\!\perp \{Y(1), Y(0)\} \qquad \text{and} \qquad \{Y(1), Y(0)\} \perp\!\!\!\perp \mathcal{X}^N. \tag{1}$$

There are two facts to notice about this setup. First, because $\mathcal{X}^S$ contains all and only the information about potential outcomes, treatment assignment Z will be independent of $\mathcal{X}^S$ if and

4. This formalization embeds the stable unit treatment value assumption (Cox 1958, Rubin 1978).

5. Assumption 1 is also sometimes called (strong) ignorability.

6. Suppose there are $\binom{N}{n_1}$ possible vectors Z in which n_1 units are assigned to treatment and n_0 units go to control. If each vector is equally likely, the chances do not depend on the vectors $\{Y(1), Y(0)\}$ so Assumption 1 holds.

7. The symbol $\perp\!\!\!\perp$ is often reserved for independence of random variables; in (1), however, $\mathcal{X}^S$, $\mathcal{X}^N$, and $\{Y(1), Y(0)\}$ are all fixed. Here, it means that $\mathcal{X}^N$ is uncorrelated with $\{Y(1), Y(0)\}$ in the finite population.

only if it is independent of potential outcomes. Second, we might observe noise covariates that are correlated with the treatment assignment process but of no relevance in predicting treatment effects; or we might observe pure noise unrelated to both treatment assignment and potential outcomes.

2.2.1 The Logic of Balance Testing

Now, denote the set of covariates that a researcher observes—i.e., actually measures—by the matrix $\boldsymbol{X}$, with rows for observations and columns for pre-treatment or placebo outcomes.

Standard practice tests the claim that $Z \perp\!\!\!\perp \boldsymbol{X}$ rather than directly testing Assumption 1. The reasoning appears to be the following:

Claim 1 *(Standard Practice: Balance tests)*

$$Z \perp\!\!\!\perp \boldsymbol{X} \iff Z \perp\!\!\!\perp \{Y(1), Y(0)\}$$

where $\iff$ means "if and only if." Hence, $Z \not\perp\!\!\!\perp \boldsymbol{X} \iff Z \not\perp\!\!\!\perp \{Y(0), Y(1)\}$.

Claim 1 is not correct, however.

Counterexample to Claim 1: False positives. Suppose that $Z \perp\!\!\!\perp \{Y(1), Y(0)\}$, so that as-if random assignment holds, and that Nature has adversarially chosen $Z \not\perp\!\!\!\perp \mathscr{X}^N$. Then if $\boldsymbol{X} \subseteq \mathscr{X}^N$, we have that $Z \not\perp\!\!\!\perp \boldsymbol{X}$ but treatment assignment is independent of potential outcomes. The $\Leftarrow$ direction of Claim 1 does not follow.

A researcher who believed Claim 1 might perform a balance test, observe imbalance between treatment and control groups on some subset of covariates, and conclude that treatment was not randomly assigned. However, this is a false positive if the imbalanced covariates are unrelated to potential outcomes: their imbalance does not constitute evidence that as-if random fails. This corresponds to one example in the introduction: men select into the treatment group in a drug trial, but gender is unrelated to potential responsiveness to treatment. We expect statistical dependence between treatment status and gender, but this does not imply that potential outcomes are imbalanced.

Conversely, a researcher might find balance on spurious covariates. But their balance does not constitute evidence that *potential outcomes* are balanced, as the next counterexample shows.

Counterexample to Claim 1: False negatives. Assume now that $Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}$, so that as-if random assignment fails, but $Z \perp\!\!\!\perp \mathscr{X}^N$. If $\boldsymbol{X} \subseteq \mathscr{X}^N$, we have $Z \perp\!\!\!\perp \boldsymbol{X}$, but it does not follow that treatment is assigned independently of potential outcomes. The $\Rightarrow$ direction of Claim 1 does not follow.

For example, suppose now that sick people select into treatment in a drug trial, yet men are as likely to do so as women—and we only measure gender. Then we would expect our balance test to produce a false negative: we would fail to reject the null but as-if random fails.

In sum, if we only measure noise covariates—those unrelated to potential outcomes—then finding balance or imbalance on those covariates does not allow us to test as-if random assignment.

2.3 The informativeness of covariates

The above discussion suggests we should consider the informativeness of covariates when constructing balance tests. We first give a sufficient condition for validly rejecting as-if random (Assumption 1) based on the non-independence of treatment assignment and covariates. For this, we use the following definition from Dawid (1979) (see also Pearl 1988; Wang and Wang 2020):

Definition 1 *([Minimal] Sufficiency of Covariates) A set of observed covariates* $\boldsymbol{X} \subset \mathscr{X}$ *is sufficient for* $Y(1), Y(0)$ *if*

$$\{Y(1), Y(0)\} \perp\!\!\!\perp \mathscr{X} | \boldsymbol{X}$$

and minimally sufficient for $Y(1), Y(0)$ *if, in addition,* $\forall \boldsymbol{S} \subset \boldsymbol{X}$*:*

$$\{Y(1), Y(0)\} \not\perp\!\!\!\perp \mathscr{X} | \boldsymbol{S}.$$

In words, if the observed covariates are sufficient for the potential outcomes, then they contain all possibly observable information about potential outcomes. Moreover, if the covariates are minimally sufficient, then they contain all *and only* the possible information (and any smaller subset $\boldsymbol{S}$ of $\boldsymbol{X}$ would no longer be sufficient).[8]

When measured covariates are minimally sufficient, standard balance tests allow us to infer whether or not Z is assigned as-if at random:

Theorem 1 *Suppose* $\boldsymbol{X}$ *is minimally sufficient for {Y(1), Y(0)}. Then,* $Z \not\perp\!\!\!\perp \boldsymbol{X} \iff Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}$.
Proof: *Appendix Section 1.*

If $\boldsymbol{X}$ is sufficient for the potential outcomes, then it must contain all the information contained in $\mathscr{X}^S$—that is, covariates that are not independent of the potential outcomes. Hence, an association between $\boldsymbol{X}$ and Z implies an association between the potential outcomes and Z. The $\Leftarrow$ direction controls false negatives: when covariates are sufficient, then when treatment is not assigned independently of potential outcomes, we should expect a well-powered balance test to fail.

If, in addition, $\boldsymbol{X}$ is *minimally* sufficient, any association between $\{Y(1), Y(0)\}$ and $\boldsymbol{X}$ will induce non-independence of $\boldsymbol{X}$ and Z. Thus, the $\Rightarrow$ direction controls false positives: when covariates are minimally sufficient, we have that a failed balance test implies a failure of as-if random.

2.3.1 Diagnosing sufficiency: the prognosis R-squared

Sufficiency could be plausible in some applied settings. For example, the lagged (pre-treatment) value of the outcome may be identically equal to the potential outcome under control—that is, $X_i = Y_{i,t-1} \equiv Y_i(0)$ for all i.[9] In this case, the pre-treatment covariate is highly *prognostic*: its correlation with potential outcomes under control is 1 in the finite population of units.

This may be why methodologists sometimes counsel the use of lagged dependent variables in placebo tests (Imbens and Rubin 2015: 483–4, Eggers, Tuñón, and Dafoe 2023, Caughey, Dafoe, and Seawright 2017). However, it is an empirical question whether and to what extent measured covariates approximate sufficient covariates—and as our empirics later in the paper show, the prognosis of covariates is often low. For lagged outcomes, the identity requires temporal stability, which would be violated if there are heterogeneous time trends in the outcome. In some applications, as we will demonstrate, pre-treatment outcomes do not in fact predict potential outcomes.

Diagnosing the joint prognosis of covariates is therefore important for evaluating the quality of tests. Goodness-of-fit measures provide one helpful tool. For example, the R^2 from the regression of $Y(0)$, as observed in the control group, on covariates indicates (linear) goodness of fit.[10] A high "prognosis R^2" indicates little residual variation in potential outcomes once we condition on covariates. Indeed, an R^2 of 1 from the finite-population regressions of both Y(0) and Y(1) on measured covariates implies sufficiency. To be sure, sufficiency could hold even with a low R^2: there

8. Equivalently, if $\boldsymbol{X}$ is sufficient, $\sigma(\mathscr{X}^S) \subseteq \sigma(\boldsymbol{X})$; moreover, if $\boldsymbol{X}$ is minimally sufficient, $\sigma(\mathscr{X}^S) = \sigma(\boldsymbol{X})$. See Lemma 1 in Appendix Section 1. This is also equivalent to Pearl 1988's notion of a Markov Blanket.

9. Units might instead all be exposed to a treatment, and a randomized intervention removes some of them. Then the logic is parallel but reversed: $Y_{i,t-1} = Y_i(1)$.

10. Later we apply this diagnostic to a sample of published studies.

may be simply be much variation in potential outcomes that is not associated with any covariates $\mathscr{X}^S$. Yet, goodness of fit makes sufficiency more plausible.

Requiring sufficiency may often set too high a bar for practice, however: it is akin to the assumption that we have measured all relevant information about potential outcomes via covariates, i.e., that we observe $\mathscr{X}^S$. Note, moreover, that Theorem 1 gives a condition for informative covariate balance tests. Yet as-if random can hold even if observed covariates are not sufficient.

The logic of Theorem 1 nonetheless suggests that even when covariates are not sufficient, we may form more specific and powerful tests by gathering covariates that are as jointly prognostic as possible. This set of jointly prognostic covariates can play the role of sufficient—but perhaps not minimally sufficient—covariates in the proof of Theorem 1. We might then further improve the performance of tests by prioritizing for testing those individual covariates in the set that are most closely associated with potential outcomes—approximating minimal sufficiency. We turn to this latter conjecture in the next two sections, where we develop our prognosis-weighted test and assess its performance.

3. An informativeness-weighted covariate balance test

The goal of our prognosis-weighted approach is to construct a test set as close to minimally sufficient as possible, via a projection of potential outcomes onto covariates. Each covariate is upweighted or downweighted according to its degree of prognosis, or the extent to which it is associated with potential outcomes. As our simulations later show, the extent to which this reduces false positives and false negatives depends on the joint prognosis of the covariates at a researcher's disposal.

3.1 A prognosis-weighted regression

Consider the conditional expectation function $\mathbb{E}(Y_i(0)|X_i)$ that gives the average value of potential outcomes under control at each value of $\boldsymbol{X}$.[11] The expectation is taken over a unit sampled at random from the finite population, or from each stratum defined by a particular value of the vector $X_i = x$.

If we observed $Y_i(0)$ for all units in the finite population, we could use linear or kernel regressions to approximate the graph of averages (the conditional expectation function). We develop our exposition using linear regression, which, as we will see, leads to a simple, attractive test with a clear connection to existing practice. The covariate-adjusted value of $Y_i(0)$ in the finite population is

$$Y_i(0)_{lr} = X_i\beta, \tag{2}$$

where "lr" denotes the linear regression and i denotes any unit. Here, X_i is a $1 \times p$ vector of covariates, and β is the $p \times 1$ vector of coefficients from the finite-population regression. Although one can define an analogous regression for $Y(1)$, we focus on $Y(0)$ because in many natural experiments, we may observe pre-treatment values of the outcome in a "no treatment" status.[12] Values of the pre-treatment outcome may tend to be especially prognostic for potential outcomes under control $Y(0)$.

Each element β_j of β is a measure of the extent to which the associated covariate is linearly prognostic. That is, it gives the "informativeness" of covariate X_j, relative also to the other covariates. Indeed, it can be represented as the coefficient from the bivariate regression

$$\beta_j = \frac{\text{Cov}(Y(0), \widetilde{X_j})}{\text{Var}(\widetilde{X_j})}, \tag{3}$$

11. Here, $\boldsymbol{X}$ is a rectangular $N \times p$ data matrix where each row is the vector X_i.
12. Mixing $Y(1)$ and $Y(0)$ values can also require additional assumptions that may not be tenable (Hansen 2008).

where $\widetilde{X_j}$ is the residual from the finite-population regression of X_j on the other $p-1$ covariates.[13] When potential outcomes and covariates are standardized by subtracting the finite-population mean and dividing by the finite-population standard deviation, each β_j is a standardized multiple regression coefficient. More prognostic covariates will have larger absolute values of standardized β_j. Conversely, the coefficient β_j vanishes when the partial correlation between $Y(0)$ and X_j is zero.

It is important to emphasize that β has no causal interpretation: the regression simply provides the best linear approximation of the potential outcomes $Y(0)$ given X in the finite population. Covariates are fixed features of units that are not here considered amenable to manipulation; even if they were, there is no expectation or requirement that manipulation would lead to expected changes in the value of the outcome variable. Nor is the correlation between $Y(0)$ and a given X_j secured as a feature of a design, such as the randomization of a treatment.

3.2 Testing as-if random

Now, define $\overline{Y(0)^T}$ as the average value of potential outcomes under control in the treatment ("T") group. Similarly, $\overline{Y(0)^C}$ is the average value of potential outcomes under control in the control ("C") group. Both are random variables when treatment assignment is randomized. Then Assumption 1 motivates the following null and alternative hypotheses:

$$\begin{aligned} H_0 &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] = 0 \\ H_A &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] \neq 0. \end{aligned} \tag{4}$$

The logic: if as-if random holds, the treatment and control group averages can be viewed as the means of samples drawn at random from the same finite population. Thus, the expected averages are the same in each sample, as under the null hypothesis H_0. Conversely, if treatment assignment were not randomized so that $Z \not\perp \{Y(1), Y(0)\}$, it would follow that the average potential outcomes in the treatment and control groups would differ in expectation, as under the alternative hypothesis H_A.

To test H_0, the problem is to estimate the unobserved difference of expectations. Consider first a regression of the outcome variable on covariates in the control group, i.e., the sample version of equation (2). We have exactly

$$\overline{Y(0)^C} = \overline{X^C}\ \widehat{\beta^C}, \tag{5}$$

where $\overline{X^C}$ (a $1 \times p$ vector) gives the average value of the p covariates in the control group and the $p \times 1$ vector $\widehat{\beta^C}$ gives the coefficients from the control group regression.[14] Descriptively, the control group regression evaluated at the average value of the covariates is $\overline{Y(0)^C}$. Moreover, equation (5) can be viewed as a regression-weighted estimator for the average potential outcome under control in the finite population: while $Y(0)_{lr}$ is incompletely observed—because we do not see $Y_i(0)$ for units in the treatment group and thus cannot fit equation (2)—the treatment and control groups are exchangeable under as-if random assignment. From another perspective, under as-if random the control group is a simple random sample from the finite population, and so we can appeal to well-known sampling theory to form a consistent estimator (e.g., Cochran 1977, Chapter 7).

We cannot run a regression analogous to equation (5) in the treatment group, however, because in that group we see potential outcomes under treatment, rather than potential outcomes under control. However, by the same logic of exchangeability, the expectation of the coefficient we would obtain—if we could run that regression in the treatment group—is clearly the same as the expectation

13. This is the Frisch–Waugh–Lovell (FWL) theorem a.k.a. "regression anatomy" (Angrist and Pischke 2009: 3.1.2)

14. Here, $\widehat{\beta^C} = (\sum_{i=1}^{n_0} X_i X_i')^{-1} \sum_{i=1}^{n_0} X_i Y_i(0)$, is a $p \times 1$ vector with elements $\widehat{\beta_j}$ for $j = 1, \ldots, p$. Here we index by $i = 1, \ldots, n_0$ the random subset of units sampled into the control group from the N units in the finite population.

of $\widehat{\beta^C}$, where the latter is viewed as a random variable. Under a null hypothesis of as-if random, we can therefore estimate the average of the potential outcomes under control in the treatment group as

$$\widehat{\overline{Y(0)^T}} = \overline{X^T}\ \widehat{\beta^C}, \tag{6}$$

where $\overline{X^T}$ is the vector of average values of covariates in the treatment group.

With an estimator of $\mathrm{E}(\overline{Y(0)^T})$ in hand, we can form a statistic to test H_0. Subtracting (5) from (6), we have

$$\begin{aligned} \widehat{\overline{Y(0)^T}} - \overline{Y(0)^C} &= (\overline{X^T} - \overline{X^C})\ \widehat{\beta^C} \\ &= \sum_{j=1}^{p} \widehat{\beta_j^C}\delta_j \\ &\equiv \delta_{PW}, \end{aligned} \tag{7}$$

with δ for "difference" and "PW" for prognosis-weighted. In (7), each δ_j is the difference of means on covariate j across the treatment and control groups—i.e., the standard test statistics in covariate-by-covariate balance tests. Each δ_j is then multiplied by the weight $\widehat{\beta_j^C}$, which is the jth coefficient from the standardized multiple regression of outcomes on covariates, as fit in the control group.

Equation (7) is our key test statistic: it is the weighted sum of individual covariate differences of means, where the weights measure prognosis for potential outcomes under control. We recommend standardizing $Y(0)$ and all covariates before running the regressions in equations (5) and (6) and forming the weighted sum δ_{PW}; this is the default option in our accompanying `R` package. This ensures that the contribution of each term to the sum is not a function of the measurement scale. In the next sub-section, we present a bootstrap hypothesis test that gives p-values for observed δ_{PW}.[15]

Use of δ_{PW} to test as-if random contrasts with standard practice in several ways. First, (i) unlike with covariate-by-covariate tests, δ_{PW} is single statistic to which we may attach a single p-value that we can use to test H_0. Next, (ii) we weight each covariate difference of means by an estimate of its informativeness. Some standard procedures—such as F-tests from the regression of treatment assignment on all covariates—avoid the problem that (i) solves. Yet, as unweighted tests they still do not prioritize prognostic covariates, a limitation that (ii) addresses. Note that rejections of as-if random in tests using δ_{PW} will be due to imbalances in covariate means, as in standard practice. Yet by upweighting prognostic covariates and downweighting non-prognostic ones the test may better detect unobserved imbalances in potential outcomes—a conjecture we assess further below. Indeed, (iii) when covariates are sufficient for $Y(0)$ in the sense of Definition 1, rejecting H_0 in a test based on δ_{PW} implies rejecting as-if random (Appendix Theorem A.3), which may not be true for standard tests. We also note that (iv) our approach gives researchers incentives to gather a wide range of covariates and include them in their tests, since this improves the prognosis R^2 and other goodness-of-fit diagnostics.

One can extend our approach in (7) to non-linear or smoothed regressions such as loess. Indeed, given data on $\boldsymbol{X}$, we could use any classification procedure, such as random forests and other machine learning techniques, to fit $Y(0)$ in the control group and use that fit to estimate $Y(0)$ in the treatment group. Such techniques may come at the cost, however, of the ready interpretation of coefficients as measures of the relative informativeness of different covariates, as in (3), though the variable importance metric in random forests is a possibility. We believe one attractive feature of our test

15. We derive the large-sample distribution of δ_{PW}, conditional on the weights $\widehat{\beta}$, in Appendix Section 2. However, the asymptotics may not apply in small studies; and the random variable $\widehat{\beta}$ is dependent on the randomness in $\overline{X^T} - \overline{X^C}$. The bootstrap test accounts for this dependence.

statistic δ_{PW} is its simplicity as well as its connection to existing practice: it is simply a sum of the standard covariate differences of means but with each difference weighted by a measure of its prognosis for potential outcomes under control. As we show by simulation later, this approach simultaneously offers gains in both sensitivity and power.

3.2.1 A bootstrap hypothesis test

Randomization tests are attractive as they allow comparison of the observed value of a test statistic to its exact randomization distribution.[16] We propose here a variant that is a resampling (a.k.a. bootstrap) technique appropriate to our setting.[17]

The procedure uses draws from the observed data to approximate the null sampling distribution of δ_{PW}, i.e., its distribution when as-if random holds. Its validity rests on three key features. (1) The expectation of the covariate difference of means in the test is zero, as it is when treatment assignment is randomized: we compare the expected values of averages of two independent samples drawn from the same finite bootstrap population.[18] (2) Critically, the procedure also allows in a natural way for the statistical dependence between the random variable $\widehat{\beta}^{C}$—as fit in the control group—and $\overline{X^{C}}$, with treatment assignment as the only source of stochastic variation. (3) Finally, the approach uses only $Y(0)$ values, rather than mixing $Y(0)$ and $Y(1)$ values as a permutation-based approach would do, which could lead to bias in the test when $\boldsymbol{X}$ predicts $Y(1)$ differently than it predicts $Y(0)$.

The resampling test works as follows, in a study with one treatment and one control group:

1. Draw a sample with replacement from the observed control group and regress outcomes on covariates. Return the coefficient vector $\widehat{\beta}^{C*}$ and the sample mean of the covariates, $\overline{X^{C*}}$.
2. Take another independent sample, also with replacement and *also from the observed control group* and calculate the sample mean of the covariates, $\overline{X^{T*}}$. Now calculate a simulated $\delta_{PW}^{*b} = (\overline{X^{T*}} - \overline{X^{C*}})'\widehat{\beta}^{C*}$.
3. Repeat steps (1)-(2) B times ($B = 500$ in our default).
4. Calculate a two-sided randomization-based p-value as

$$p^{*} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\delta_{PW}^{*b}| \geq |\delta_{PW}^{\text{obs}}|), \tag{8}$$

   where $\mathbb{I}$ is an indicator function that takes on the value of 1 if its argument is true and 0 otherwise. Reject the null if, e.g., $p^{*} < 0.05$.

This bootstrap procedure can be adapted to accommodate a wide range of designs, for instance, those with clustered or blocked assignment. We also note that using control group values to estimate the weights does not induce a bias from overfitting, a problem that can arise when study outcomes are also used for estimating average treatment effects (Rubin 2007; Hansen 2008; Liao et al. 2023).

3.2.2 Equivalence testing

The testing procedure can also be adapted to take advantage of equivalence tests (Hartman and Hidalgo 2018). In our context, the null hypothesis would be that as-if random does *not* hold—i.e.,

16. See Fisher (1935); also inter alia Caughey, Dafoe, and Seawright (2017).
17. Freedman 2009: Section 8.1.
18. The observed treatment and control group means are dependent and the samples are drawn without replacement. However, X_i is the same whether unit i is assigned to treatment or control. Per Neyman (1923), it is thus as if the two samples were drawn independently with replacement (see Freedman, Pisani, and Purves 2007: A32-A34; Samii and Aronow 2012, Theorem 2; Gerber and Green 2012: 57; or Dunning 2012: 193).

the data are inconsistent with an unconfounded design—while the alternative says that the expected values of the average potential outcomes in the two groups are approximately equal:

$$\begin{aligned} H_0 &: \mathbb{E}[\overline{Y(0)}^T - \overline{Y(0)}^C] \neq 0 \\ H_A &: \mathbb{E}[\overline{Y(0)}^T - \overline{Y(0)}^C] \approx 0. \end{aligned} \tag{9}$$

What constitutes approximate ($\approx$) or sufficient equality under the alternative is captured by the "equivalence range," the requirement that $E[\overline{Y(0)}^T - \overline{Y(0)}^C]$ be contained in some interval $[\epsilon_L, \epsilon_U]$.

Equivalence testing seeks to address the "balance test fallacy" (Imai, King, and Stuart 2008, Section 7), in particular, the problem that failing to reject the null of as-if random is not the same as accepting it: researchers may fail to reject simply because their study is small and underpowered. Equivalence tests are less likely to reject the null of difference as a study grows smaller (e.g., as control group observations are randomly dropped, see Hartman and Hidalgo 2018, Figure SI-2). One drawback is the need to define an equivalence range, which introduces discretion: researchers can find evidence for or against as-if random by varying the range. Alternatives that lessen this discretion—for instance, use of the equivalence confidence interval (Hartman and Hidalgo 2018)—make equivalence testing more akin to traditional balance testing since in the latter, one can also readily examine a $(1-\alpha)*100\%$ confidence interval to see what parameter values lie outside of it.

The best advice may be to develop high-powered tests—either traditional or equivalence-based—with prognostic covariates and carefully consider the substantive range of true differences that are consistent with the intervals. Our prognostic-weighted statistic facilitates this since, under either an equivalence or traditional framework, the test is based on the balance of the most prognostic covariates.

3.3 Regression-discontinuity designs: testing the continuity of potential outcomes

Our procedure can also be adapted to RD designs, including the close-election designs we will study later. In some RD designs, the assumption of as-if random within a small bandwidth around the threshold determining treatment assignment may be the relevant condition to test, in which case the techniques we have developed so far straightforwardly apply.[19]

However, analysts have rightly noted that in other RD settings, as-if random should be replaced with the (weaker) assumption that regression functions relating potential outcomes to the forcing variable (a.k.a., the "score") are continuous at the threshold determining treatment assignment (see e.g. Calonico, Cattaneo, and Titiunik 2014 and De la Cuesta and Imai 2016). Continuity implies that the limits of these functions approaching the threshold from above and below are the same. For $Y(0)$, for example,

$$\lim_{r \downarrow c} \mathbb{E}[Y_i(0)|R_i = r] = \lim_{r \uparrow c} \mathbb{E}[Y_i(0)|R_i = r]. \tag{10}$$

Here, R_i is the value of the forcing variable, and c is the threshold value of R_i at which treatment Z_i switches "on" or "off". The expectations operators refer to the expected value for a randomly sampled unit with $R_i = r$. Informally, the "intercepts" of two regressions of potential outcomes on the forcing variable—one above and another below the discontinuity—must be the same.

Yet, perhaps because some potential outcomes are unobservable, researchers typically test for the continuity not of $Y(0)$—as in equation (10)—but of *covariates*. Thus, they regress each placebo or pre-treatment covariate separately on the forcing variable, above and below the RD threshold. For each placebo/covariate, they then test a null hypothesis that the intercepts of the two regressions (one above and one below) are equal at the threshold.

19. This may especially be so when the slope of the regression function relating potential outcomes to the forcing variable is flat (see Dunning 2012 Chapters 3 and 5; Cattaneo, Frandsen, and Titiunik 2015; Sekhon and Titiunik 2017).

However, such tests for the continuity of covariates may not be informative about the continuity of *potential outcomes*. Just as with tests of as-if random, researchers are subject to false negatives and false positives due to irrelevant covariates (Theorem 1). Potential outcomes may be continuous at the threshold and yet covariates may not be, or vice versa.

Fortunately, we can readily form a prognosis-weighted test statistic that is appropriate for testing continuity of potential outcomes in RD designs. Following our previous approach of using only the prognostic part of the covariates, let $\widehat{Y(0)} = \boldsymbol{X}\widehat{\beta^C}$ be the fitted value from a regression of the outcome on covariates on the control group side of the RD threshold, where $Y(0)$ is observed. The test for continuity of potential outcomes is then (mirroring e.g. De la Cuesta and Imai 2016: 385-6) based on

$$\delta_{PW}^{RD} \equiv \widehat{\alpha_1} - \widehat{\alpha_0}. \tag{11}$$

Here,

$$(\widehat{\alpha_0},\ \widehat{\beta_0}) = \underset{\alpha_0,\,\beta_0}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c_0 \leq R_i \leq c\}\{\widehat{Y_i(0)} - \alpha_0 - \beta_0(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{12}$$

is the intercept and slope from a regression of $\widehat{Y(0)}$ on the forcing variable to the right of the threshold (centered at the threshold). Similarly,

$$(\widehat{\alpha_1},\ \widehat{\beta_1}) = \underset{\alpha_1,\,\beta_1}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c < R_i \leq c_1\}\{\widehat{Y_i(0)} - \alpha_1 - \beta_1(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{13}$$

is the intercept and slope from the regression to the right. For clarity, we separate the fitted intercepts $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$ from $\widehat{\beta_0}$ and $\widehat{\beta_1}$, the fitted coefficients on the forcing variable R_i. Note the latter are distinct from the fitted coefficients of the regression of $Y(0)$ on *covariates* $\boldsymbol{X}$, which we label $\widehat{\beta^C}$ as before.

Thus, we form the test statistic in (11) as the difference of prognosis-weighted intercepts of regressions above and below the threshold. Here, $\widehat{\alpha_1}$ and $\widehat{\alpha_0}$ are $\widehat{Y(0)}|R_i = c$, i.e., the predicted potential outcomes under control at the threshold, conditional on the covariates to the right and to the left of the threshold, respectively. Conceptually, it is as if we were separately regressing each pre-treatment covariate on the forcing variable in the windows $R_i \in [c_0, c]$ and $R_i \in [c, c_1]$ below and above the RD threshold c, as in standard practice, but we then combine the intercepts of these separate regression lines into prognosis-weighted intercepts from each side of the RD threshold.[20] One can readily adapt the approach analogously to test the continuity of $Y(1)$, though again, in some applications $\boldsymbol{X}$ may be most prognostic for $Y(0)$ (e.g., when covariates include a lagged outcome).

We can then test the following null hypothesis against the alternative:

$$\begin{aligned} H_0 &: \mathbb{E}[\widehat{\alpha_1} - \widehat{\alpha_0}] = 0 \\ H_A &: \mathbb{E}[\widehat{\alpha_1} - \widehat{\alpha_0}] \neq 0, \end{aligned} \tag{14}$$

or we can flip the null and alternative hypotheses, as in equivalence testing. Appendix Section 3.2 discusses further details.

As with as-if random, this test of continuity projects out irrelevant covariates and thus bases assessment on the covariates associated with potential outcomes. When covariates are jointly prognostic, the approach allows for a more powerful and specific test of the continuity of *potential outcomes*—rather than covariates.

20. As recommended by Calonico, Cattaneo, and Titiunik 2014 and Cattaneo, Idrobo, and Titiunik 2020, equations (12) and (13) are triangular kernel-weighted local linear regressions; $K(\cdot)$ may be a function such as the triangular kernel, $K(u) = (1 - |u|) \cdot \mathbb{I}\{|u| < 1\}$. The bandwidth $[c_0,\ c_1]$ can be chosen by the algorithm of Imbens and Kalyanaraman 2012; this is the default option in our `R` package `pwtest`.

4. Simulations

The extent to which prognosis weighting boosts the power and specificity of balance tests may vary across different data sets and data-generating processes, which makes the tests' performance well-suited for investigation via simulations. We conduct a large set of simulations that allow us to study rejection rates of our test when as-if random holds and when it is false, varying covariate prognosis. When as-if random is false, the rejection rate measures statistical power of the test; whereas when it is true, the proportion of rejections measures the false positive rate (or Type I error, inversely related to specificity). We also compare the performance of the test to two unweighted multivariate test statistics, the unweighted sum of standardized covariate differences of means and Hotelling's T^2, which do not make use of information about covariate prognosis. Due to space limitations, we present simulation results in the online Appendix (Section 4).

The simulations illustrate that by projecting out irrelevant covariates, prognosis weighting can reduce both false negatives and false positives. In contrast, unweighted tests that do not use information on covariate prognosis sacrifice power and/or specificity. The results offer an important caveat, however: consistent with our theoretical results, the quality of tests—including prognosis-weighted ones—depends on the overall joint prognosis of measured covariates.

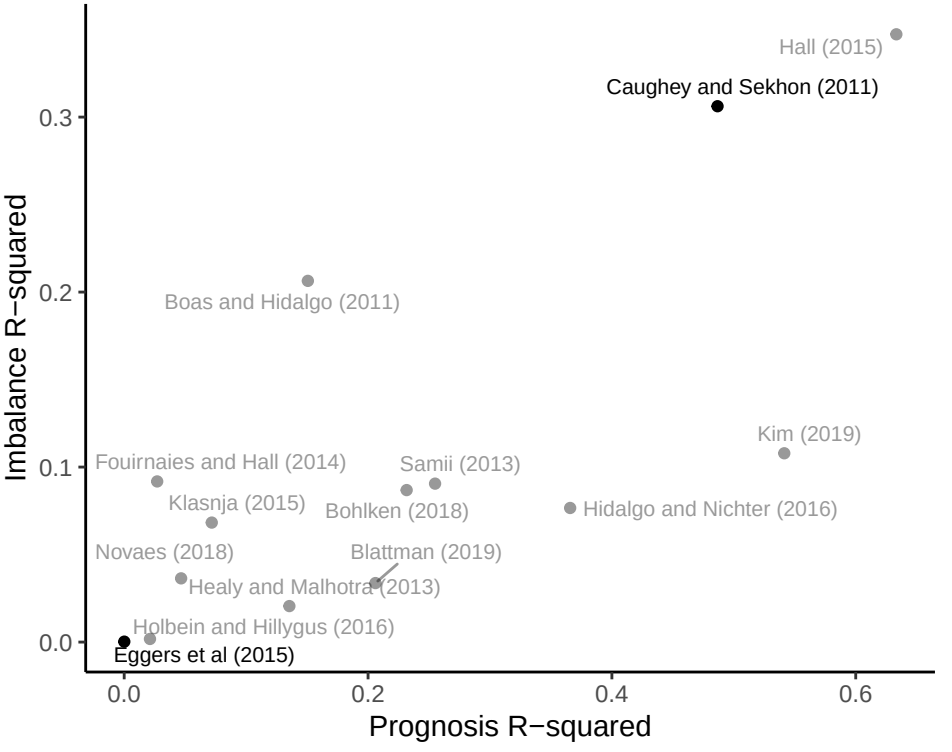
5. The varied prognosis of covariates

Covariate prognosis is rarely considered systematically in balance tests, despite its importance. We coded a random sample of 150 articles that use randomized experiments, natural experiments, and RD designs in three top political science journals (the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*), stratifying by journal, over the time period 2000-2019.[21] Overall, 52 percent of the sampled articles presented covariate balance tests in the paper or Appendix. The majority of those (56 percent, or 29 percent overall) use only covariate-by-covariate tests, rather than an omnibus test statistic; those that present the latter typically report the p-value of the F-statistic for the regression of a treatment indicator on all covariates.[22] Only 18 percent of tests used a lagged dependent variable as a covariate. And we found no examples of systematic efforts to account for the informativeness of covariates in balance tests.

The covariates used in balance tests vary substantially, nonetheless, in the extent to which they predict potential outcomes. We took a small further random sample from the 150 studies, excluding randomized experiments, stratifying by natural experiment versus discontinuity and on the presence of a lagged dependent variable or explicit discussion of prognosis in the paper. We had to exclude some natural experiments either due to lack of appropriate replication data or other considerations (Appendix Section 5.1). We calculated two measures for each included study: the multiple R^2 from the regression of observed potential outcomes in the control group on all available covariates ("Prognosis R^2") and the multiple R^2 from the regression of a treatment assignment indicator on all available covariates ("Imbalance R^2").

Figure 1, which plots these measures for our smaller sample of studies, suggests several insights. First, there is considerable variation across studies in the extent to which the covariates used in balance tests are prognostic. Close to the vertical axis, covariates are non-prognostic and thus bear little apparent relationship to potential outcomes. Moving towards the right side of Figure 1, however, we have several studies in which covariates are more associated with potential outcomes. Second, in this sample of published natural experiments and discontinuity designs, overall we find relatively little imbalance. Thus, most of the studies cluster along the bottom part of the plot, where the R^2 for the regression of treatment assignment on covariates is low. This is perhaps natural: studies in which covariates are substantially imbalanced are unlikely to be published as natural experiments. Still, there

21. For code used in the sampling, see OMITTED.
22. See Hansen and Bowers 2008 on drawbacks of this procedure.



The figure plots a sample of natural experiments and regression-discontinuity (RD) designs drawn from all those published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*, 2000-2019; Caughey and Sekhon (2011) is added. Prognosis R^2 comes from a regression of potential outcomes under control on all available covariates (control group only). Imbalance R^2 comes from a regression of treatment assignment on all available covariates. Two studies we discuss in detail later are bolded. See Appendix Section 5 for further information.

Figure 1. Imbalance vs. Prognosis In Balance Testing (Sample of Natural Experiments and RD Designs)

are studies in which the imbalance R^2 is relatively high. Sampling a fuller range of observational studies would presumably populate the top half of the figure to a greater extent.

As our discussion so far suggests, this variation in prognosis—as well as imbalance—can be critical for purposes of assessing as-if random or continuity of potential outcomes. Heuristically, there are four kinds of cases in Figure 1. (1) Studies located in the upper-left quadrant may be prone to spurious rejection with standard procedures—because there is imbalance on covariates unrelated to potential outcomes. (2) In the lower-left quadrant, the important concern is instead that none of the measured covariates are prognostic of potential outcomes but we find balance on treatment assignment—leading to a form of Type I error in which we fail to reject as-if random, yet potential outcomes themselves may be imbalanced. (3) In the lower-right quadrant, we find cases with high prognosis but low imbalance: here, the claim of as-if random may be most persuasive. (4) Finally, in the upper-right quadrant, rejection may be most persuasive of a failure of as-if random—because covariates are as a whole prognostic of potential outcomes. We note, however, that location in this quadrant need not imply rejection when one uses the prognosis-weighted procedure: covariates may be associated with potential outcomes as a whole, leading to a high prognosis R^2, and yet imbalance may occur on a non-prognostic subset of covariates.

5.1 Prognosis weighting in practice

Figure 2 reports prognosis-weighted and unweighted tests of as-if random for the full sample of studies using natural experiments or RD designs (depicted in Figure 1). For those studies using discontinuities, we also include results of our test for continuity of potential outcomes. We fail to reject as-if random (or continuity of potential outcomes, for RD studies) using any test in 5 of these 14 these papers. In other papers, however, a prognosis-weighted test rejects where an unweighted test does not, or vice versa.

Figure 3 lends insight into why this divergence occurs. Here, we plot the difference of means associated with each covariate in each study against each covariate's prognosis (standardized regression coefficient). Inspection of the results for each study suggests that where highly prognostic covariates are imbalanced but non-prognostic covariates are balanced, the weighted test rejects but not the unweighted test. When instead it is the prognostic covariates that are balanced, the opposite appropriately occurs. In other words, the test results are weighted towards the most prognostic covariates.

5.2 Case study: are close elections random?

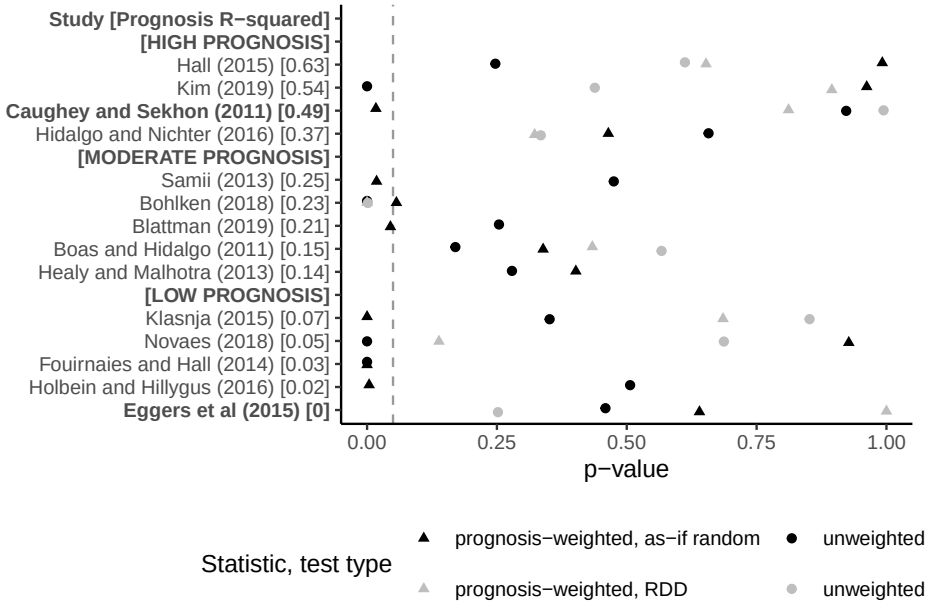
A controversy over the randomness of close elections illustrates the importance of covariate prognosis in balance tests, as well as the value of an omnibus informativeness-weighted test.

A priori, in a very close election, which party winds up with a slightly greater vote share at time t may seem quite plausibly as-if random (Lee 2008, Lee and Lemieux 2010). If so, assignment of party incumbency is independent of potential outcomes, facilitating study of the impact on electoral outcomes at time $t + 1$. For this reason, the close-election design has become extremely widespread.

In an important study, however, Caughey and Sekhon (2011) critically appraise the assumption of as-if random in close U.S. House elections (1942-2008). Caughey and Sekhon present a series of covariate difference-of-means tests in a small neighborhood around the vote threshold determining party incumbency, i.e., around a 0% difference in Democrat-Republican vote share. Concerningly, p-values from balance tests suggest statistically significant imbalances in past incumbency, as well as the winning party's past vote share, campaign spending, and measures of candidate quality. Perhaps party incumbency is therefore not as-if randomly assigned after all in very close elections.[23]

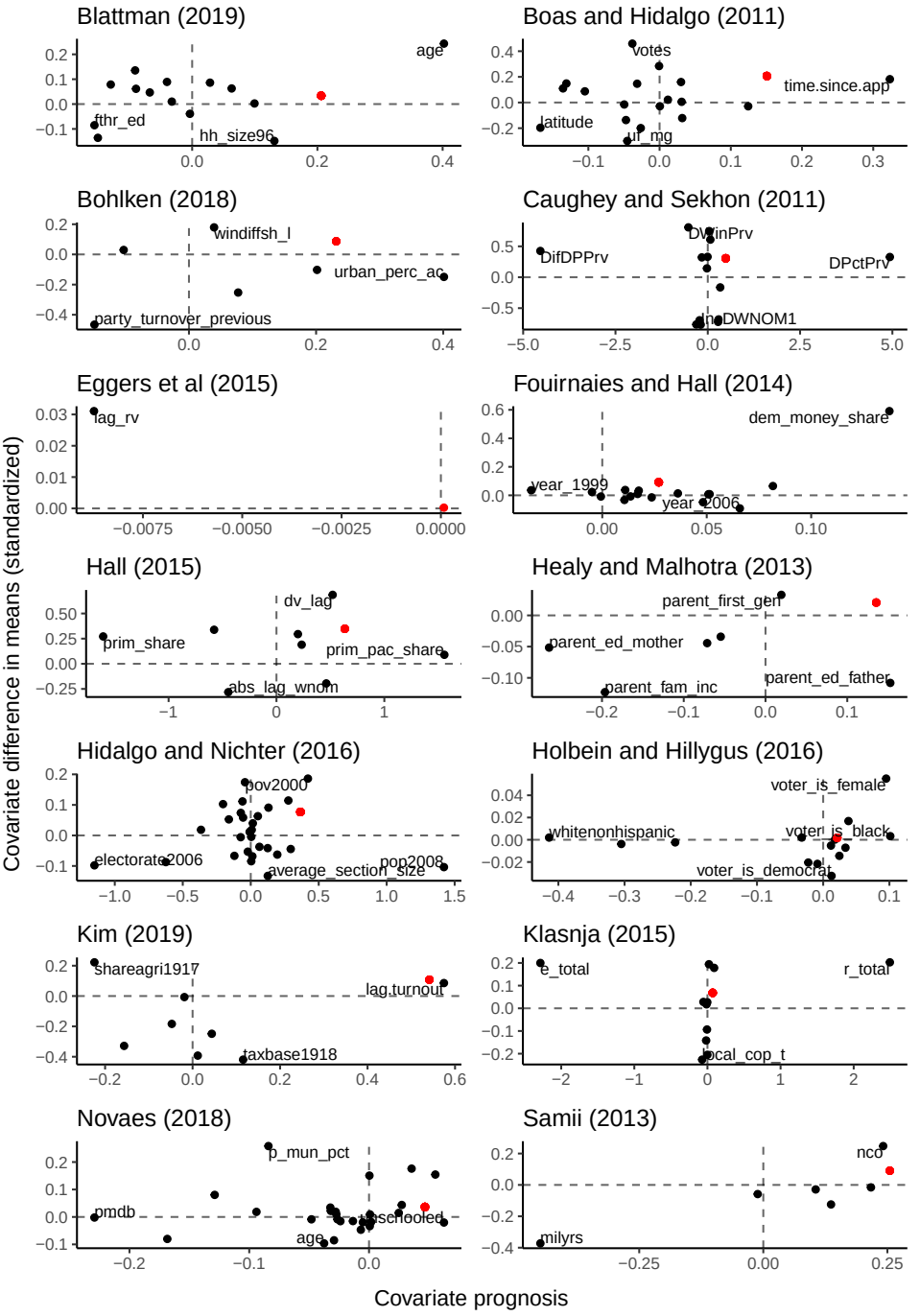
In an excellent subsequent study, Eggers et al. 2015 challenge this conclusion. They confirm that lagged incumbency seems to be the major driver of imbalances in Caughey and Sekhon's data: in a

23. See their Figure 2. Caughey and Sekhon (2011) also present evidence of "sorting" at the 0% vote margin.



The figure plots p-values for prognosis-weighted (triangles) and unweighted (circles) tests, for the sample of studies in Figure 1. The number in [brackets] is the prognosis R^2. For RD studies, we include p-values from tests of continuity of potential outcomes (light shading). The dotted vertical line is at $p = 0.05$.

Figure 2. Prognosis weighting in practice



For each of our sampled studies (see Figure 2), here we plot for each covariate the standardized difference of means across treatment and control (vertical axis) against the covariate's standardized multiple regression coefficient, from the prognosis regression (horizontal axis). The red dots indicate the overall prognosis and imbalance R^2s.

Figure 3. Unpacking the weights

procedure conceptually related to ours, they show that Democratic near-winners and near-losers are not significantly different on pre-treatment covariates other than lagged incumbency, once the latter is controlled in a regression. They then extend the Caughey and Sekhon study to a broad range of majoritarian elections around the world, comparing close election winners and losers on a measure of lagged incumbency. The authors find balance on lagged incumbency in every other close-election setting they examine. They therefore suggest the observed imbalance in U.S. House elections may reflect special features of that context or is simply be due to chance.

Returning to the original U.S. House data, De la Cuesta and Imai (2016) also show that the evidence against key assumptions in that context may be weaker than it had appeared. They convincingly argue that in this context—where the forcing variable, Democratic vote share, may have a strong relationship to future Democratic incumbency—researchers should test not as-if random but instead continuity of potential outcomes at the RD threshold. Using kernel-weighted local-linear regressions like our equations (12) and (13) instead of differences of means, and also correcting for multiple testing concerns, they show much weaker treatment-control imbalances than Caughey and Sekhon.[24]

What does this controversy reveal about the importance of covariate prognosis, and what are implications of the specific procedure we propose? In Figure 2, we report *p*-values from prognosis-weighted as well as unweighted tests applied to data from Caughey and Sekhon (2011) and Eggers et al. (2015) (in bold), as well as the sample of natural experiments presented in Figure 1 (discussed next).

We can draw several conclusions. First, our approach addresses important concerns about multiple testing. In Caughey and Sekhon 2011, as in many studies, tests based on different specific covariates give rise to different conclusions: while variables such as lagged incumbency are imbalanced, districts barely won by Democrats at time t do not differ from those barely won by Republicans on several political and demographic variables.[25] Moreover, covariates are correlated, so simply comparing the number of rejections to the number of tests (e.g. to see if the ratio is greater than 1 out of 20) is not informative. As De la Cuesta and Imai (2016) rightly emphasize, multiple statistical comparisons can also lead to false positives (Benjamini and Hochberg 1995). Our omnibus approach provides a new way to elide such multiple testing concerns while accounting for dependence of the covariates: rejection can be based on a single test statistic combining information from all the covariates.

Second, formally accounting for the relative prognosis of different covariates, as in our approach, allows conclusions based on variables most related to potential outcomes. This can be seen with the test of as-if random using the Caughey and Sekhon data. Here, an unweighted test does not in fact reject as-if random (dark circle in Figure 2). However, an omnibus prognosis-weighted test—one that projects out irrelevant covariates unrelated to potential outcomes—does (dark triangle in Figure 2). Figure 3, which we discuss later, shows large imbalances among two highly prognostic variables in these data. This accords with Caughey and Sekhon's view that variables such as lagged incumbency are particularly critical for testing. However, unlike existing approaches, our prognosis weighting allows formal incorporation of relative covariate informativeness in the balance tests.

Third, our adaptation of prognosis weighting to RD designs also allows a test of continuity of potential outcomes. Figure 2 shows that our prognosis-weighted test based on δ_{PW}^{RD} in equation (11) does not reject the weaker assumption of continuity in the Caughey and Sekhon 2011 data, even though the test is based on the most prognostic covariates. Thus, while differences of means suggest imbalances on informative covariates, consistent with De la Cuesta and Imai (2016) we do not find evidence against continuity when upweighting those prognostic variables.

24. See also Hartman 2021, who compares results with the Caughey and Sekhon (2011) data using equivalence and traditional tests.

25. For example, Caughey and Sekhon find balance on whether the state has a Democratic governor or secretary of state; the margin of victory in the presidential race; voter turnout; whether the seat is open; and the percentage of urban, Black, or foreign-born residents.

Finally and perhaps most importantly, the controversy underscores the critical role of joint prognosis of the covariates overall—and the need to diagnose and analyze it formally. As noted, Caughey and Sekhon attribute particular importance to the imbalance on the winning party's past incumbency and vote share at time $t-1$. In the U.S. House, their set of covariates is indeed jointly prognostic (Figure 1). However, they do not measure the prognosis of the lagged dependent variables or formally take it into account in their tests. Eggers et al., building on this idea, effectively assert that lagged incumbency is the only important covariate on which to test for balance—since it is the sole pre-treatment covariate on which they test for balance in their global dataset.[26] This approach is entirely understandable as well as practical: a measure of lagged party incumbency is readily available across elections and countries, whereas the availability of other pre-treatment covariates may vary by context. However, these studies and many like them do not assess the prognosis of covariates empirically or incorporate that information into their balance tests.

In fact, the prognostic value of lagged incumbency in close election designs varies across countries and types of elections. As we show in Tables A4-A5 in Appendix Section 5.2, in the Eggers et al. data, the correlation between the vote share of the incumbent party at time $t-1$ and time t is 0.79 across all countries and election types but varies from a low of 0.09 in Brazilian mayoral elections to a high of 0.91 in the German Bundestag (full data set); in close elections (defined by a bandwidth of 0.5, i.e., the margin between the winning and runner-up party is less than 1 percentage point), it varies from a high of 0.32 in New Zealand's post-war parliament to a low of –0.16 is the Canadian House of Commons (1867-1911).[27]

Perhaps most importantly, the average prognosis is essentially zero across all close elections studied by Eggers et al (Appendix Table A5 and Figure 1). It is substantially higher in the post-war U.S. House elections studied by Caughey and Sekhon (prognosis R^2 of 0.83 in the full data and 0.49 in close elections; Figure 1).[28] With Eggers et al.'s cross-national dataset, the prognosis-weighted tests reject neither as-if random nor continuity (Figure 2). Yet the low prognosis of covariates raises the risk of false negatives, per our theoretical and simulation results. Moreover, since these data include lagged incumbency as the only covariate, a weighted test cannot effectively partial out the relative prognosis of different covariates, boosting specificity and power.

We thus do not view our evidence as supporting or invalidating the key assumptions of close election designs in general: we suspect they might hold in some contexts and not in others. Yet, the results instead suggest the importance for testing of leveraging a richer set of informative covariates across global contexts. Overall, the case study thus underscores the critical relevance of diagnosing prognosis and of prioritizing covariates most associated with potential outcomes.

6. Conclusion

Applied researchers should seek to test observable implications of assumptions that facilitate causal inference. Yet this paper shows that a widely used technique that purports to accomplish this goal may not, in fact, do so.

We demonstrate that prognosis is a critical consideration for balance testing. Yet, it receives little formal or diagnostic attention. To assess the association between treatment assignment and potential outcomes, researchers should use covariates jointly associated with potential outcomes. This helps to avoid both false negatives and false positives. They should report diagnostic measures, such as the

26. Eggers et al. (2015: 262-3) argue that (a) the variety of characteristics on which winners and losers of close elections may vary can all be viewed as proxies for (are highly correlated with) incumbency; (b) testing for other covariates introduces multiple testing concerns; and (c) incumbency "confers electoral benefits in a variety of electoral settings around the world."

27. See Schiumerini (2015) on the varied effects of incumbency across contexts.

28. Restricting the analysis to close elections may attenuate correlations by truncating the range of variation on incumbent vote share at time t; yet this is the relevant subset of the data in which to assess prognosis, since this is the set in which balance tests are typically conducted.

prognosis R^2. Finally, they should prioritize the most individually prognostic covariates, as in our informativeness-weighted test.

We view our contribution as a foundation for future extensions and refinements. One complementary direction may be the use of non-linear approximations to $E(Y_i|X)$, including various machine learning techniques, as well as regressions that make greater use of information on $Y(1)$. As we have described, however, such approaches may have downsides relative to our approach. The test statistics we propose are simple and relate readily to the tests used in standard practice. The prognosis weights are interpretable in terms of the relative informativeness of different covariates.

Testing itself is complementary to other objectives, including optimization of observed balance in observational studies. Our topic also has connections to sensitivity analysis (Rosenbaum 2010). Yet, testing has an important independent role, particularly for design-based analysis of natural experiments. As-if random—if true—allows simple and credible estimation of treatment effects (Freedman 2009). We have thus focused on improving tests of whether identifying assumptions are true—rather than on strategies for limiting bias when they fail.

Combining diverse qualitative and quantitative evidence may best allow assessment of key assumptions that facilitate causal inference (Dunning 2012). In natural experiments and related designs, however, we have shown that the observed balance of covariates across treatment and control groups can be irrelevant—when they are not associated with potential outcomes. By instead leveraging the power of prognosis, researchers can build more useful, informative tests.

References

Angrist, Joshua D., and Jors-Steffen Pischke. 2009. *Mostly harmless econometrics: an empiricist's companion.* Princeton University Press.

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. ISSN: 00359246, accessed May 6, 2024. http://www.jstor.org/stable/2346101.

———. 1997. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* 24 (3): 407–18.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82 (6): 2295–2326. ISSN: 00129682, 14680262, accessed October 5, 2023. http://www.jstor.org/stable/43616914.

Cattaneo, Matias D., Brigham R. Frandsen, and Rocío Titiunik. 2015. Randomization inference in the regression discontinuity design: an application to party advantages in the u.s. senate. *Journal of Causal Inference* 3 (1): 1–24.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2020. *A practical introduction to regression discontinuity designs: foundations.* Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press. https://doi.org/10.1017/9781108684606.

Caughey, Devin, Allan Dafoe, and Jason Seawright. 2017. Nonparametric combination (npc): a framework for testing elaborate theories. *The Journal of Politics* 79 (2): 688–701. https://doi.org/10.1086/689287.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. Elections and the regression discontinuity design: lessons from close u.s. house races, 1942-2008. *Political Analysis* 19 (4): 385–408. ISSN: 10471987, 14764989, accessed May 2, 2022. http://www.jstor.org/stable/41403727.

Cochran, William G. 1977. *Sampling techniques.* John Wiley & Sons.

Dawid, A. P. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41 (1): 1–31. ISSN: 00359246. http://www.jstor.org/stable/2984718.

De la Cuesta, Brandon, and Kosuke Imai. 2016. Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science* 19 (1): 375–396.

Dunning, Thad. 2008. Improving causal inference: strengths and limitations of natural experiments. *Political Research Quarterly* 61 (2): 282–293. ISSN: 10659129, accessed October 30, 2023. http://www.jstor.org/stable/20299732.

———. 2012. *Natural experiments in the social sciences: a design-based approach.* Strategies for Social Inquiry. Cambridge University Press. https://doi.org/10.1017/CBO9781139084444.

Eggers, Andrew C., Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Jr. Snyder. 2015. On the validity of the regression discontinuity design for estimating electoral effects: new evidence from over 40,000 close races. *American Journal of Political Science* 59 (1): 259–274.

Eggers, Andrew C., Guadalupe Tuñón, and Allan Dafoe. 2023. Placebo tests for causal inference. *American Journal of Political Science* n/a (n/a): n/a. https://doi.org/https://doi.org/10.1111/ajps.12818. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12818. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12818.

Fisher, Ronald A. 1935. *The design of experiments.* Oliver & Boyd.

Freedman, David A. 1999. From association to causation: some remarks on the history of statistics. *Statistical Science* 14 (3): 243–258. https://doi.org/10.1214/ss/1009212409. https://doi.org/10.1214/ss/1009212409.

———. 2009. *Statistical models: theory and practice.* Cambridge University Press.

Freedman, David A., Robert Pisani, and Roger Purves. 2007. *Statistics.* Fourth. W.W.Norton.

Gagnon-Bartsch, Johann, and Yotam Shem-Tov. 2019. The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics* 13 (3): 1464–1483. https://doi.org/10.1214/19-AOAS1241. https://doi.org/10.1214/19-AOAS1241.

Genovese, Christopher R., Kathryn Roeder, and Larry Wasserman. 2006. False discovery control with p-value weighting. *Biometrika* 93 (3): 509–24.

Gerber, Alan S., and Donald P. Green. 2012. *Field experiments: design, analysis, and interpretation.* W.W. Norton & Co.

Hansen, Ben B. 2008. The prognostic analogue of the propensity score. *Biometrika* 95 (2): 481–488.

Hansen, Ben B., and Jake Bowers. 2008. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 23 (2): 219–236. https://doi.org/10.1214/08-STS254. https://doi.org/10.1214/08-STS254.

Hartman, Erin. 2021. Equivalence testing for regression discontinuity designs. *Political Analysis* 29 (4): 505–521. https://doi.org/10.1017/pan.2020.43.

Hartman, Erin, and F. Daniel Hidalgo. 2018. An equivalence approach to balance and placebo tests. *American Journal of Political Science* 62 (4): 1000–1013. https://doi.org/https://doi.org/10.1111/ajps.12387. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12387. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12387.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81 (396): 945–960. ISSN: 01621459. http://www.jstor.org/stable/2289064.

Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6 (2): 65–70.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502. https://doi.org/https://doi.org/10.1111/j.1467-985X.2007.00527.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-985X.2007.00527.x. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2007.00527.x.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: an introduction.* Cambridge University Press. https://doi.org/10.1017/CBO9781139025751.

Imbens, GW, and K. Kalyanaraman. 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies* 79 (January): 933–959. https://doi.org/10.1093/restud/rdr043.

Kost, James T., and Michael P. McDermott. 2002. Combining dependent p-values. *Statistics & Probability Letters* 60 (2): 183–190. https://EconPapers.repec.org/RePEc:eee:stapro:v:60:y:2002:i:2:p:183-190.

Leacy, FP, and EA Stuart. 2014. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med* 33 (20): 3488–508. https://doi.org/10.1002/sim.6030.

Lee, David S. 2008. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics* 142 (2): 675–697.

Lee, David S., and Thomas Lemieux. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* 48, no. 2 (June): 281–355. https://doi.org/10.1257/jel.48.2.281. https://www.aeaweb.org/articles?id=10.1257/jel.48.2.281.

Liao, Lauren D., Yeyi Zhu, Amanda L. Ngo, Rana F. Chehab, and Samuel D. Pimentel. 2023. *Using joint variable importance plots to prioritize variables in assessing the impact of glyburide on adverse birth outcomes.* https://doi.org/10.48550/ARXIV.2301.09754. https://arxiv.org/abs/2301.09754.

Liu, Keli, and Feng Ruan. 2020. *A self-penalizing objective function for scalable interaction detection.* arXiv: 2011.12215 [stat.ME].

Neyman, Jerzy Splawa, D. M. Dabrowska, and T. P. Speed. 1923. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (Translated 1990). *Statistical Science* 5 (4): 465–472. https://doi.org/10.1214/ss/1177012031. https://doi.org/10.1214/ss/1177012031.

Pearl, Judea. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1558604790.

Rosenbaum, Paul. 2010. *Design of observational studies.* Springer Series in Statistics, January. ISBN: 978-1-4419-1212-1. https://doi.org/10.1007/978-1-4419-1213-8.

Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66 (5): 688–701.

———. 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine* 26:20–36.

Rubin, Donald B., and Neal Thomas. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95 (450): 573–585. https://doi.org/10.1080/01621459.2000.10474233. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474233. https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474233.

Samii, Cyrus, and Peter Aronow. 2012. On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters - STAT PROBAB LETT* 82 (February). https://doi.org/10.1016/j.spl.2011.10.024.

Schiumerini, Luis Enrique. 2015. Incumbency and democracy in south america. PhD diss., Yale University.

Sekhon, Jasjeet S., and Rocío Titiunik. 2017. On interpreting the regression discontinuity design as a local experiment. *Advances in Econometrics* 38:1–28.

Stuart, EA, BK Lee, and FP Leacy. 2013. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol* 66 (8): S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013.

Wainstein, Leonard. 2022. Targeted function balancing. UCLA Department of Statistics.

Wang, Yue, and Linbo Wang. 2020. Causal inference in degenerate systems: an impossibility result. In *Proceedings of the twenty third international conference on artificial intelligence and statistics,* edited by Silvia Chiappa and Roberto Calandra, 108:3383–3392. Proceedings of Machine Learning Research. PMLR, 26–28 Aug. https://proceedings.mlr.press/v108/wang20i.html.

Westfall, Peter H. 2014. *Combining p-values.* John Wiley / Sons, Ltd.