ONLINE APPENDIX

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

Clara Bicalho*, Adam Bouyamourn†, and Thad Dunning‡

October 28, 2025

*Tinker Postdoctoral Fellow, Stanford University
†Postdoctoral fellow, Department of Politics, Princeton University
‡Professor of Political Science, Department of Political Science, University of California, Berkeley.

Contents

1 Covariate prognosis in published balance tests

1.1 Sample of experiments, natural experiments, and discontinuity designs

We searched for articles containing the keywords "randomized experiment", "natural experiment" and "regression discontinuity design" in their abstract or main body published in the three top journals in political science (the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*) between 2000 and 2020. From the results of this query, we randomly sampled 150 articles, stratifying by journal.[1]

We manually reviewed the sample of 150 studies and coded the type of design and whether they included balance tests either in the main text or in supplementary materials. We then further restricted our sample to studies that were either (a) natural experiments or regression discontinuity designs (excluding randomized control trials) and (b) that included balance or placebo tests.[2] The final sample contained 40 studies. Our analysis includes all 16 out of these 40 studies for which replication data was available and complete.

Below, we describe in more detail our processing of the data for each study in our final sample and highlight analytical choices we made when needed (e.g., specifying which of the outcomes or bandwidths our analysis used when authors' analysis included multiple outcomes or bandwidths). We also note which studies were excluded from the sample and why.

1.1.1 Included studies

Blattman (2009)

Blattman (2009) studies the relationship between violence and political participation of ex-combatants. The paper rests on the assumption that abduction into the Lord's Resistance Army (and consequent experience of violence) was exogenous, though conditionally on the age of children. We focus our analysis on the first outcome studied: whether an individual voted in 2005, and we use the covariates for which the author reports balance tests in Table 2. We restrict the analysis to interviewed subjects only, so that the prognosis and balance analyses use the same sample.

Boas and Hidalgo (2011)
In the first part of their study, Hidalgo and Boas (2011) use a regression discontinuity design to measure the effect of incumbency on politician's control of local media. The authors use raw vote margin as a forcing variable. We use the experimental sample defined by the optimal bandwidth determined by the authors, which is of 165 votes. Our analysis includes all 18 covariates the authors report in their placebo test in Table 1.

Eggers et al. (2015) Eggers et al. (2015) use a regression discontinuity design of close elections to measure incumbency advantage in competitive elections. Authors perform this analysis on elections data in different countries. We use the lagged running variable as a covariate (vote margin of reference party in

[1] For code used in the sampling, see https://github.com/[ANONYMIZED]/JSTOR_query.

[2] Keywords sometimes returned studies that cited natural experiments, but were not employing the design, for example. In all cases, we coded the design according to authors' own labelling of their study as a natural experiment or regression discontinuity design.

time $t-1$, vote margin in time t as an outcome, and incumbency at time t as a treatment. Our analysis restricts the sample to .5% margins around the cutoff point (i.e., the "naïve" specification) (p. 264, fn 12). Using a difference of means within this bandwidth is similar to the approach in Caughey and Sekhon (2011)'s paper on the US House and thus facilitates comparisons of findings in the two papers.

Fouirnaies and Hall (2014)
Fouirnaies and Hall (2014) use a regression discontinuity design to estimate the effect of incumbency on campaign contributions in the U.S. House and state legislatures. The authors run separate regressions on samples of state and federal legislature. We randomly select the state sample to use in our analysis, further restricting the data to the sample of vote margin $\leq 1\%$, the smallest bandwidth used by the authors (this uses the same sample as Table 1, column 4). In our analysis, we include covariates used in Table 4 of the Appendix: Democratic Party's share of contributions in election t, year dummies, state dummies and a chamber dummy (for the state legislatures).

Hall (2015)
Hall uses a regression discontinuity design to estimate the effects of extremist candidates winning primaries on the party's vote share in the general elections. There are three outcomes of interest: party vote share, party victory, and the DW-NOMINATE score of the winning general election candidate in the ensuing Congress. In our analysis, we use the first outcome: party vote share. We also use covariates used in the author's results on covariate balance on Table A5: "These variables are as follows: absolute distance from 50% of the presidential normal vote in the district (the Democratic presidential normal vote for Democratic primaries, and the Republican presidential normal vote for Republican primaries), averaged over the period 1980–2010, to measure the partisanship of the district; the extreme candidate's share of primary donations; the extreme candidate's share of primary donations from PACs; the absolute value of the district's previous incumbent's DWNOMINATE score, to measure the ideology of the district; the absolute value of the district's previous incumbent's WNOMINATE score; and the party's lagged vote share and electoral victory" (p. 35). It is worth noting that the authors use year fixed effects in one of their balance tests (variable "abs_lag_wnom"), but not the others. In our model of prognosis, we have excluded year fixed effects. In footnote 42, the author argues "the balance tests turn out to be highly similar without these year fixed effects." We use a 0.05% vote margin in the as-if random tests, following the sample the author uses.

Healy and Malhotra (2013)
Healy and Malhotra (2013) are interested in the effect of socialization (in particular having sisters) on the attitude changes among men. They use the gender of the younger sibling as an instrument for share of a respondent's siblings who are female. Our analysis focuses on the first set of results using the Political Socialization Panel (PSP) survey. There are three survey waves of PSP (1973, 1982, 1997). We randomly picked one wave: 1973. Authors also use two specifications: whether number of siblings is included as linear controls or as fixed effect. We former the latter specification in our definition of the covariate matrix. We use the instrument as the treatment variable. We define covariates according to the balance tests reported in SI Figure S1.

Hidalgo and Nichter (2016)
Hidalgo and Nichter (2016) exploit a discontinuity in audit probability to examine the effect of vote

buying (which is undermined by audits) on reelection rates of mayors in Brazil. We process the data following the authors' replication files, including imputing missing values with median of non-missing values. We restrict the RD sample according to the optimal bandwidth authors use in the difference in means analysis (1.5%) in the percentage of electorate as a share of total population in 2006. Our analysis also focuses on the first outcome: change in voter registration between 2007 and 2008. It is worth noting that the authors' balance test analysis in Figure 4 includes 'electoratechange0708.perpop' as a covariate, although that variable is the same as the one used as the outcome variable in the authors' analysis. The label in Figure 4 suggests that the covariate refers to electorate change between 2002 and 2007. However, the code book does not include a reference to the latter. Instead, in our analysis we use 'electorate-change0407.perpop' as a covariate, which is the change in electorate between 2004 and June 2007 and as a percentage of 2007 population. We also exclude 'num_vereadores04' (the number of local councilors in 2004) because the covariate values are constant in the non-missing control observations and we are unable to estimate the variance of our unweighted test statistics or the prognosis coefficient of that covariate.

Holbein and Hillygus (2016)
We produce statistics for the "Analysis 2: Florida Voter File" portion of Holbein and Hillygus (2016). Authors explore a fuzzy regression discontinuity design by using date of birth cutoff for eligibility to vote to measure the effect of preregistration on voter turnout around the elibility cutoff. We use the control variables the authors include in their balance test in Table A2 and the treatment variable as defined by the cutoff point (eligibility to vote). The bandwidth of 18 days for the as-if random test is the same used by authors in Table A2. The test of continuity uses eligibility to vote as treatment (reduced form) and the bandwidth used by authors to generate results in Table 3 (36 days on either side of the cutoff).

Kim (2019)
Kim (2019) exploits a discontinuity design based on a population threshold that assigns direct democracy to municipalities in Sweden to study the effect of direct democracy on the political inclusion of newly enfranchised women. In our analysis, we use the first (and apparent primary) outcome analyzed in the paper: women's turnout. Although the author's analysis pools outcome data across multiple years (with year fixed effects), ours uses outcome data from 1921 (the first post-treatment year observed in the data set) as we believe it is the best sample on which to assess prognosis. Our analysis includes all covariates for which the author tests balance in Figures SI 1.1-1.3: left parties' vote share in 1917, turnout in 1917, ENEP in 1917, share of organized citizens, tax base income in 1918, percentage of the agriculture in the economy in 1917, land area in 1918, and number of poor relief participants in 1917. We exclude percentage of female attendees in municipal meetings in 1917 since it is only available for a subset of the data. The bandwidth we use mimics the optimal bandwidth used by Kim of approximately 286 people around the cutoff population size. This choice of bandwidth is used in both the test of as-if random and of continuity.

Klašnja (2015) Our analysis focuses on Klašnja (2015)'s regression discontinuity design to examine incumbency advantage in Romanian mayoral elections. We use the continuous outcome, vote margin, in our analysis, and define the sample using the optimal bandwidth reported on Table 1, column (2) ($bw = 0.149$). The covariates we use are the ones included on the balance table in the Appendix (Table A5). Whereas authors use a continuous treatment (vote margin) as an instrument for incumbency in the main results, our analysis uses a binary variable—win_t—as the treatment indicating incumbency status. This is because

there is one-sided non-compliance— not all observations for whom vote margin is greater than 0 correspond to elected officials due to the way the authors choose to code run-off elections (see discussion in the study's Appendix Section A2).

Novaes (2018)
We draw from the balance tests reported in the Supplementary Information, Table 2 which includes 27 covariates. For our imbalance analysis, we define the bandwidth at 0.5% (the fourth column in Table 2, same bandwidth we use in other RDs such as Caughey and Sekhon (2011) and Eggers et al. (2015)). For prognosis, there are two main outcome variables in the paper (p. 89): party switching (by the candidate) and party electoral performance (vote shares for congressional candidates in the winning or losing brokers). We use the former in our analysis, since covariates are also measured at the candidate level and we consider it a more relevant as a measure of prognosis.

Samii (2013)
We run our analysis on the regression discontinuity sample in the study (bandwidth of 5 years above and below the threshold). We use the covariates that comprise the author's first placebo test in Table 4 (columns 1-5). According to Samii (2013): "As a further robustness check, I conduct "placebo" tests with variables that could not possibly have been causally affected by treatment (Imbens and Lemieux 2008). One wants to do this on pretreatment variables that have strong potential to confound were they to exhibit discontinuities near the cutoff." These variables are non-commissioned officer status, years in the military, years of education pre-war, unit death rate, and family death rate.[3] The specifications in the paper involve a two-stage least squares analysis using location above or below the age threshold for service in an ethnically integrated military as an instrumental variable for actual integration. Since what is proposed as as-if random in this natural experiment is location above or below the age threshold for service of 45 years (within the 5-year bandwidth), we conduct covariate balance tests using this location (i.e. the value of the instrument) to define treatment and control groups.

Thomas (2018)
Thomas (2018) measures partisan bias in the allocation of public resources. The author uses a regression discontinuity design relying on close races to evaluate the impact of co-partisanship between local MPs and state legislators on the allocation of development project proposals in India. The author tests balance on a set of six covariates in Table A3. These variables are Margin of Victory of MLA in Previous Term, Party Turnover of MLA in Previous Term, Percent Literacy, Percent Urban, Percent SC/ST, and Percent Agricultural Laborers. As the author describes "the estimates [in Table A3] are obtained by estimating Equation 1 with the relevant pre-treatment covariate as the dependent variable. The key independent variable is Co-Partisan State Incumbent. Each specification includes a quadratic polynomial in the Forcing variable and an interaction of each of these terms with the variable Co-Partisan State Incumbent. Controls include the variables Urban, Allotment Increase and Multiple. State fixed effects, project year fixed effects and parliament fixed effects are also included." In our analysis, we include only the six pre-treatment covariates the author considers in the balance test. We adopt the bandwidth of other close-election study designs we consider (0.5% margin) for the tests of as-if random.

[3]Samii (2013: 569-70) also includes a measure item nonresponse for nonsensitive questions in his survey as a placebo outcome. We do not include that post-treatment variable in our analysis in Figure 2 in the paper.

1.1.2 Excluded studies

Arceneaux, Lindstädt, and Wielen (2016): Arceneux et al. examine the effect of partisan news media on legislative behavior. They exploit the incremental rollout of Fox News Channel in the late 1990 to compare legislative behavior among Democrats and Republicans across districts without Fox News and districts with partial Fox News access. Authors perform a balance test and results are reported in Table 1, where results on covariate balance are reported for the following covariates: whether legislator is a Democrat, ideology, seniority, 1996 spending gap, 1996 challenger spending, 1996 quality challenger, 1996 incumbent wins, and 1996 presidential vote. We were not able to identify these covariates in the data. The authors' replication material does not include the script used to generate Table 1, and in the absence of a guidebook, we cannot be sure of which data columns refer to which covariates.

Branton et al. (2015): The authors refer to their design as a natural experiment but did not include a covariate balance test in the main text or supplementary material, so we regarded the replication materials as incomplete.

Enns and Richman (2013): This study proposes to measure the effect of election salience (measured by voters receiving voter guide on state elections) on voters' incentive to accurately report their presidential vote incentive. Authors argue their study is a natural experiment. For the first part of the analysis, they compare outcomes across different windows of the survey period and show that outcomes differ from zero for a specific window which coincides with a time when all subjects received treatment. Treatment is administered to all registered voters in California at the same point in time (no randomization), and these voters are compared with voters in other states. Authors use CEM matching to units outside of the treated state to account for confounders, but the original set up does not rely on the "as-if" random assumption. The second part of the analysis compares phone and in-person interviews, assuming subjects are randomly sampled from the broader US population. This approach more closely resembles a natural experiment. However, there is no "control" group per se. Rather, because the comparison is between phone and in-person surveys it is difficult to justify our approach of using the "control" sample to measure prognosis — more specifically, the justification of using the control group to measure covariate prognosis to the entire sample in expectation is not well adjusted to this research setting.

Galasso and Nannicini (2011): This paper doesn't actually use the balance test in our usual sense. It proposes a theory of political selection whereby parties nominate different types of political candidates in safe elections but converge on the same type of "high quality" candidate in close elections. Then the authors test for balance in close elections but that is understood as the absence of partisan differences in candidate characteristics (i.e. differences across the parties in candidate types in the close elections). Because difference among politicians on either sides of the cutoff are treated as an outcome, rather than a placebo test for the effect of induced by the cutoff, we decided to exclude this study from our analysis.

Longo, Canetti, and Hite-Rubin (2014): The data provided with the replication materials did not contain covariates authors used in the balance checks script (Table 1). These variables were "extremism", "Religion_Ideology", and "Religion_Behavior" not included in the data provided.

Malesky, Nguyen, and Tran (2014): We could not find the data file that contains the full set of covariates that the authors test balance on (46 covariates in Table 1) and excluded this study from our replication.

Velez and Newman (2019): Authors had to suppress a key variable of their data set required for their analysis for privacy purposes, so we were not able to perform our estimation.

In addition, we could not find replication data in the public domain for the following studies: Schickler, Pearson, and Feinstein (2010), Ferwerda and Miller (2014), Chauchard (2014), Davenport (2015), Kam and Palmer (2008), Hirano (2011), Shami (2012), McClurg (2006), Eggers and Hainmueller (2009), Findley, Nielson, Sharman (2015), and Mendelberg, McCabe, and Thal (2017).[4]

1.2 Close elections: the prognosis of lagged incumbency

In this subsection, we analyze the prognosis of lagged incumbency across country-election types, using the data from Eggers et al. (2015). Table A1 uses all the data, while Table A2 restricts the analysis to the RD study group with bandwidth 0.5, i.e., those elections where the margin of victory between the two leading parties is less than 1 percent. The final column of each table shows the correlation between Y_0—the party vote share at time t in the control group–and X, the party vote share at time $t-1$.

Table A1: Eggers et al.: Correlation between party vote shares at times t and t-1 across election types (all)

Country	Office	Corr_Y0_X
USA	HOUSE OF REPRESENTATIVES, 1880-2010	0.852
USA	HOUSE OF REPRESENTATIVES, 1880-1944	0.8712
USA	HOUSE OF REPRESENTATIVES, 1946-2010	0.8339
USA	STATEWIDE	0.7465
USA	STATE LEGISLATURE	0.7681
USA	MAYOR	0.6029
CANADA	COMMONS, 1867-2011	0.7457
CANADA	COMMONS, 1867-1911	0.5383
CANADA	COMMONS, 1921-2011	0.7569
UK	HOUSE OF COMMONS	0.8434
UK	LOCAL COUNCIL	0.7883
GERMANY	BUNDESTAG	0.9139
GERMANY	BAVARIA, MAYOR	0.4255
FRANCE	NATIONAL ASSEMBLY	0.7019
FRANCE	MUNICIPALITY	0.6667
AUSTRALIA	HOUSE OF REPS, 1987-2007	0.904
NEW ZEALAND	PARLIAMENT, 1949-1987	0.8043
INDIA	LOWER HOUSE, 1977–2004	0.4316
BRAZIL	MAYORS, 2000-2008	0.0847
MEXICO	MAYORS, 1970-2009	0.7465
All COUNTRIES	**ALL RACES**	**0.7907**

[4]Mendelberg, McCabe, and Thal (2017) did provide replication materials, but these included only scripts and "read me" files and no data.

Table A2: Eggers et al.: Correlation between party vote shares at times t and t-1 across election types (RD study group with bandwidth 0.5—close winners and close losers)

Country	Office	Corr_Y0_X
USA	HOUSE OF REPRESENTATIVES, 1880-2010	0.1417
USA	HOUSE OF REPRESENTATIVES, 1880-1944	0.0649
USA	HOUSE OF REPRESENTATIVES, 1946-2010	0.2389
USA	STATEWIDE	-0.1045
USA	STATE LEGISLATURE	4e-04
USA	MAYOR	0.0173
CANADA	COMMONS, 1867-2011	-0.064
CANADA	COMMONS, 1867-1911	-0.1625
CANADA	COMMONS, 1921-2011	-0.0383
UK	HOUSE OF COMMONS	0.1764
UK	LOCAL COUNCIL	0.0513
GERMANY	BUNDESTAG	-0.052
GERMANY	BAVARIA, MAYOR	-0.1254
FRANCE	NATIONAL ASSEMBLY	-0.0647
FRANCE	MUNICIPALITY	0.1305
AUSTRALIA	HOUSE OF REPS, 1987-2007	0.2946
NEW ZEALAND	PARLIAMENT, 1949-1987	0.3236
INDIA	LOWER HOUSE, 1977–2004	-0.073
BRAZIL	MAYORS, 2000-2008	-0.0487
MEXICO	MAYORS, 1970-2009	0.0065
All COUNTRIES	**ALL RACES**	**0.0221**

Note that the overall average in the final row of Table A5—the average correlation across all country-election types—is 0.02, whereas it is essentially zero in Figure 1 in the paper. The very minor difference stems from the use here of the 0.5 bandwidth—the preferred bandwidth of Eggers et al. (2015)—whereas in Figure 1, we use the MSE-optimal (default) bandwidth returned by the `R` routine `rdrobust`, via our package `pwtest` in which rdrobust is a dependency (Calonico et al. 2015). In this case, the default returns a bandwidth of about 0.415 rather than 0.5. See our replication code for Tables A1 and A2 for further details.

2 The logic of standard balance tests

In this section, we develop the argument made in Section 3 of the paper formally, using a design-based, finite population set-up. We use the same notation and framework for exposition of our prognosis-weighted tests in section 7 of the paper and section 7.2 in this appendix.

Thus, here we show why balance tests based on the independence of treatment assignment and covariates typically do not validly test independence of treatment and potential outcomes, give a condition under which they do, and discuss measurement and reporting of covariate prognosis.

2.1 Defining as-if random

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.[5] 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 $n_1 + n_0 = N$. The set-up is design-based in that the only source of random variation is the treatment assignment vector Z; potential outcomes are fixed.

In valid natural experiments, the following condition must hold:

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

where $\perp\!\!\!\perp$ means "is independent of."[6] In words, the random treatment variable is assigned independently of potential outcomes.[7] As-if random ensures, for example, that sicker patients do not go systematically to the treatment group in a drug trial studying health outcomes, or that those more 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).

The assumption of as-if random, however, can be the "Achilles Heel" of natural experiments (Dunning 2008). In a true randomized experiment, a chance protocol under the control of a researcher (Fisher 1935) ensures that treatment is independent of potential outcomes, as well as any fixed covariates. In natural experiments, by contrast, 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. While qualitative information and ancillary tests may support this assertion (Dunning 2012), 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).

With additional assumptions discussed below, however, it is possible to make as-if random falsifiable.

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

[6] Assumption 1 is also sometimes called (strong) ignorability.

[7] 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.

2.2 Standard balance tests, and two counterexamples

When can the balance of observed covariates test as-if random? To answer this question, consider a space of possible covariates $\mathscr{X}$. Suppose that $\mathscr{X} = \{\mathscr{X}^S, \mathscr{X}^N\}$ ('signal' and 'noise', respectively – compare Liu and Ruan 2020), where $\mathscr{X}^S$ contains all information about potential outcomes and $\mathscr{X}^N$ contains none. Treat $\mathscr{X}^S$ and $\mathscr{X}^N$ as finite but potentially unobserved. Thus,

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

where "$\not\perp\!\!\!\perp$" means "not independent of."

There are two facts to notice about this setup. First, because $\mathscr{X}^S$ contains all and only the information about potential outcomes, treatment assignment Z will be independent of $\mathscr{X}^S$ if and 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 the units and columns containing measured pre-treatment covariates or placebo outcomes. We emphasize that $\boldsymbol{X}$ may not coincide with the possible covariates $\mathscr{X}$: the problem is that researchers may only be able to collect data on a subset of these possible covariates. The measured covariates may thus contain some, all, or none of the signal variables $\mathscr{X}^S$.

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.

For example, in an observational study of the efficacy of a new drug, men might tend to select into the treatment group. Yet gender may be unrelated to health status or responsiveness to the treatment. If we have only data on gender, we may wrongly reject as-if random based on the covariate imbalance, even though potential outcomes themselves may be balanced in expectation.

Conversely—and perhaps most importantly, as we might worry most about false claims to a natural experiment—balance on a spurious covariate does not imply that treatment is assigned independently of potential outcomes, 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, sicker patients might select into the treatment group. As-if random may thus fail. Health after an intervention may be closely related to prior health, yet we may fail to measure the latter, signal covariate. In contrast, men may be as likely to select into treatment as women, leading to expected balance on gender. Yet, if gender is not related to potential outcomes or responsiveness to treatment, its observed balance cannot readily validate as-if random. If we base a balance test on gender, we may thus falsely fail to reject as-if random.

In sum, covariates differ in their informativeness about potential outcomes. 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.

3 The importance of covariate prognosis for balance tests

Here we give a condition for validly rejecting as-if random (Assumption 1 in the text and in section 2 above), based on the non-independence of treatment assignment and covariates.

3.1 Sufficiency of 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*

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

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

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

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]

[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. This is also equivalent to Pearl (1988)'s notion of a Markov Blanket.

3.1.1 Minimal sufficiency and balance tests (Theorem A.1, statement and proof)

When measured covariates are minimally sufficient, as-if random fails if and only if treatment assignment depends on covariates:

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:*** *See subsection 3.1.1.*

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: if covariates are minimally sufficient, a failed balance test implies a failure of as-if random.

To prove Theorem A.1, we first develop alternate definitions of sufficiency and minimal sufficiency, showing that these are equivalent to the definition above.

Suppose that there exists a unique $\mathscr{X}^S \subseteq \mathscr{X}$ such that $\sigma(\mathscr{X}^S) = \sigma(\{Y(1), Y(0)\})$, where $\sigma(\cdot)$ denotes sigma algebras. Then we have:

Definition A.1 *Minimal sufficiency of covariates (alternate version of Definition 1 in the text)*

$$\text{If } \boldsymbol{X} \text{ is sufficient, } \sigma(\mathscr{X}^S) \subseteq \sigma(\boldsymbol{X}); \text{ moreover,}$$

$$\text{If } \boldsymbol{X} \text{ is minimally sufficient, } \sigma(\mathscr{X}^S) = \sigma(\boldsymbol{X}).$$

Note that these definitions of sufficiency and minimal sufficiency are equivalent to those given above, as the next lemma shows.

Lemma 1. *Definition 1 holds* $\iff$ *Definition A.1 holds.*

For sufficiency, take any $\boldsymbol{X'} \subseteq \mathscr{X} \setminus \boldsymbol{X}$. Then

$$\begin{aligned}
\sigma(\mathscr{X}^S) \subseteq \sigma(\boldsymbol{X}) &\iff \sigma(\{Y(1), Y(0)\}) \subseteq \sigma(\boldsymbol{X}) \\
&\iff \sigma(\boldsymbol{X'} \cap \{Y(0), Y(1)\}) = \emptyset, \quad \forall \boldsymbol{X'} \quad \text{(by the definition of } \boldsymbol{X'}) \\
&\iff \{Y(1), Y(0)\} \perp\!\!\!\perp \boldsymbol{X'} \\
&\iff \{Y(1), Y(0)\} \perp\!\!\!\perp \mathscr{X} | \boldsymbol{X}.
\end{aligned}$$

The idea is that, when $\boldsymbol{X}$ is sufficient, there can be no other variable $\boldsymbol{X'}$ that also contains information about the potential outcomes, in which case potential outcomes are conditionally independent of any other variable given $\boldsymbol{X}$. Conversely, if $\boldsymbol{X}$ is not sufficient, there must be such an $\boldsymbol{X'}$, and we do not have the conditional independence of potential outcomes and $\mathscr{X}$ given $\boldsymbol{X}$.

For minimal sufficiency, we start with Definition A.1, which implies that $\boldsymbol{X}$ is minimally sufficient if, in addition: $\forall \boldsymbol{S} \subset \boldsymbol{X},\ \exists \boldsymbol{X'} \subset \mathscr{X} \setminus \boldsymbol{S}$, such that $\{Y(1), Y(0)\} \not\perp\!\!\!\perp \boldsymbol{X'}|\boldsymbol{S}$. (This says that there must be some subset of $\mathscr{X} \setminus \boldsymbol{S}$ that contains information about potential outcomes.) Then,

$$\begin{aligned}
\sigma(\mathscr{X}^S) = \sigma(\boldsymbol{X}) &\iff \sigma(\{Y(0), Y(1)\}) = \sigma(\boldsymbol{X}) \\
&\iff \forall \boldsymbol{S} \subset \boldsymbol{X}, \sigma(\boldsymbol{S}) \subset \sigma(\boldsymbol{X}) = \sigma(\{Y(0), Y(1)\}) \\
&\iff \exists\, \boldsymbol{X'} \subseteq \mathscr{X} \setminus \boldsymbol{S} \text{ s.t. } \sigma(\boldsymbol{X'} \cap \{Y(0), Y(1)\}) \subseteq \sigma(\{Y(1), Y(0)\}) \\
&\iff \{Y(1), Y(0)\} \not\perp\!\!\!\perp \boldsymbol{X'}|\boldsymbol{S} \\
&\iff \{Y(1), Y(0)\} \not\perp\!\!\!\perp \mathscr{X}|\boldsymbol{S}
\end{aligned}$$

Intuitively, if $\boldsymbol{X}$ is minimally sufficient, then any strict subset of $\boldsymbol{X}$ does not include all information about potential outcomes; so there is some set in $\boldsymbol{X}$ not in the smaller set that also has information about potential outcomes. Therefore, conditioning on the smaller set does not make the collection of possible covariates conditionally independent of potential outcomes. (Compare Pearl 1988). Thus, Definition 1 can hold if and only if Definition A.1 holds.

Proof of Theorem A.1: With these preliminaries, we can prove Theorem A.1, which states:

> Assume $\boldsymbol{X}$ is minimally sufficient for {Y(1), Y(0)}. Then, $Z \not\perp\!\!\!\perp \boldsymbol{X} \Leftrightarrow Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}$.

Proof. We show the $\Rightarrow$ direction by contrapositive, noting that

$$\begin{aligned}
(\neg Z \not\perp\!\!\!\perp \{Y(1), Y(0)\} &\implies \neg Z \not\perp\!\!\!\perp \boldsymbol{X}) \\
&\iff \\
(Z \not\perp\!\!\!\perp \boldsymbol{X} &\implies Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}).
\end{aligned}$$

$$\begin{aligned}
\neg Z \not\perp\!\!\!\perp \{Y(1), Y(0)\} &\implies Z \perp\!\!\!\perp \{Y(1), Y(0)\} \\
&\implies \sigma(Z) \perp\!\!\!\perp \sigma(\{Y(1), Y(0)\}) \\
&\implies \sigma(Z) \perp\!\!\!\perp \sigma(\boldsymbol{X}) && \text{[By minimal sufficiency]} \\
&\implies Z \perp\!\!\!\perp \boldsymbol{X} \\
&\implies \neg Z \not\perp\!\!\!\perp \boldsymbol{X}
\end{aligned}$$

We have verified the contrapositive, which allows us to conclude that $Z \not\perp\!\!\!\perp \boldsymbol{X} \implies Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}$.

To show the ($\Leftarrow$) direction, note that assuming $\boldsymbol{X}$ is minimal sufficient implies that $\boldsymbol{X}$ is sufficient. Then:

$$\begin{aligned}
Z \not\perp\!\!\!\perp \{Y(1), Y(0)\} &\implies \sigma(Z) \not\perp\!\!\!\perp \sigma(\{Y(1), Y(0)\}) \\
&\implies \sigma(Z) \not\perp\!\!\!\perp \sigma(\mathscr{X}^S) \\
&\implies \sigma(Z) \not\perp\!\!\!\perp \sigma(\boldsymbol{X}) && \text{[By sufficiency]} \\
&\implies Z \not\perp\!\!\!\perp \boldsymbol{X}
\end{aligned}$$

□

In sum, we have assumed the existence of a set $\mathscr{X}^S$ that must contain all the information in the potential outcomes. If covariates $\boldsymbol{X}$ are minimally sufficient they must contain all *and only* the information in the potential outcomes. Hence constructing a test of the independence of Z and a minimal sufficient covariate set $\boldsymbol{X}$ is equivalent to constructing a test of the independence of Z and the potential outcomes.

4 Prognosis-weighted tests of as-if random

In this section, we describe the theoretical distributions of several statistics used in the paper and provide formal details on the bootstrap hypothesis test of as-if random.

4.1 The fitted value approach

Consider as before a study with a 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 under exposure to treatment and to control, 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.[9] 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. This set-up is design-based in that the only source of random variation is the treatment assignment vector Z; potential outcomes are fixed.

As-if random (Assumption 1) motivates the following testable null and alternative hypotheses:

$$\begin{aligned} H_0 &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] = 0 \\ H_A &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] \neq 0. \end{aligned} \tag{2}$$

Here, $\overline{Y(0)^T}$ is the average value of potential outcomes under control in the treatment ("T") group sample, while $\overline{Y(0)^C}$ is the average value of potential outcomes under control in the control ("C") group sample. Both are random variables when treatment assignment is randomized.

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 $Z \not\perp\!\!\!\perp \{Y(1), Y(0)\}$, 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. This in turn requires a procedure for predicting $\overline{Y(0)^T}$ in the treatment sample, where potential outcomes under control are not observed. With a procedure for forming this fitted value in hand, we can form a test statistic as the difference

$$\delta_{PW} = \widehat{\overline{Y(0)^T}} - \widehat{\overline{Y(0)^C}}, \tag{3}$$

that is, the fitted average $Y(0)$ in the treatment group minus the fitted average $Y(0)$ in the control group. Essentially, we fit $Y(0)$ in the control group, which gives us prognosis weights, and then apply this weighting

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

procedure to the covariates in the treatment group. Thus, δ_{PW} is the prognosis-weighted ("PW") difference ("δ") in fitted values across the treatment and control groups. We focus on control group regressions, as in e.g. Hansen (2008) and Stuart et al. (2013), because pre-treatment values of the outcome variable, which are sometimes measured, may tend to be especially prognostic for potential outcomes under control.

4.2 A regression-based test

Consider the sample regression of the outcome variable on covariates in the control group:

$$\begin{aligned} \widehat{\overline{Y(0)^C}} &= \overline{X^C}\,\widehat{\beta^C} \\ &= \overline{Y(0)^C}, \end{aligned} \tag{4}$$

where the $1\times p$ vector $\overline{X^C}$ 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.[10] Descriptively, the control group regression evaluated at the average value of the covariates is exactly the sample average $\overline{Y(0)^C}$.

While we cannot fit the analogous finite-population regression—because we do not see $Y_i(0)$ for units in the treatment group—under as-if random the control group is a simple random sample from the finite population. Equation (4) can thus be viewed as a regression-weighted estimator for the average potential outcome under control in the finite population (Cochran 1977, Chapter 7).

We cannot run a regression like equation (4) in the treatment group, because in that sample we see potential outcomes under treatment, rather than potential outcomes under control. However, by the same logic, the expectation of the coefficient we would obtain—if we could run that regression in the treatment sample—is clearly the same as the expectation 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

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

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

Subtracting (4) from (5) gives an estimator of the unobserved difference of the expectations (3), valid under H_0. Thus we have

$$\begin{aligned} \widehat{\mathrm{E}}[\overline{Y(0)^T} - \overline{Y(0)^C}] &= (\overline{X^T} - \overline{X^C})\,\widehat{\beta^C} \\ &= \sum_{j=1}^{p} \widehat{\beta_j^C}\,\delta_j \\ &\equiv \delta_{PWLR}, \end{aligned} \tag{6}$$

with "PWLR" for "prognosis-weighted linear regression."

In (6), each δ_j is the difference of means on covariate j across the treatment and control groups. The weight $\widehat{\beta_j^C}$ is the jth coefficient from the multiple regression of outcomes on covariates in the control group. We recommend standardizing $Y(0)$ and all covariates before forming the fitted values in equations (4) and (5) and the weighted sum in (6); this is the default option in our accompanying `R` package. This ensures

[10] 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.

that the contribution of each term to the sum is not a function of the measurement scale. The standardized regression coefficients will be larger in absolute value for more prognostic covariates, while they vanish when the partial correlation between $Y(0)$ and X_j is zero.

It is important to emphasize that β, the coefficient of the finite- population regression corresponding to (4), has no causal interpretation: the regression simply provides the best linear approximation of the potential outcomes $Y(0)$ given X. 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. The procedure simply allows for measurement of covariate prognosis.

The use of the test statistic δ_{PWLR} has several possible advantages, relative to those we consider next. One is its simplicity and intelligibility: the weights (regression coefficients) are readily interpretable as the relative informativeness or prognosis of the respective covariate, relative also to the other covariates.[11] Another is its close connection to current practice. Rejection of as-if random in tests using δ_{PW} will be due to treatment-control differences of covariate means, as in standard covariate-by-covariate tests (Section 2). Yet, unlike standard practice, the test prioritizes the variables most informative about potential outcomes, potentially allowing more powerful and specific tests—a conjecture we evaluate in section **??**. Finally, as with other procedures we consider next, the test combines information on prognosis across covariates to form an ombnibus statistic to which we may attach a single p-value to test H_0.

4.3 Theoretical properties of unweighted and prognosis-weighted tests

Here we discuss several theoretical properties of test statistics presented in the paper or used in the simulations. First, we derive the theoretical distribution of the unweighted sum of covariate differences of means, which is useful as a benchmark since our key test statistic δ_{PWLR} is the weighted version of this sum. We then discuss its relationship to Hotelling's T^2, another unweighted test used for testing multivariate equality of means, and we relate this to the F-distribution. In our simulations, we compare the performance of tests based on the unweighted sum and on Hotelling's T^2 to those based on key test statistic δ_{PWLR}.

While we also derive the large-sample conditional distribution of our key test statistic δ_{PWLR}, conditional on the weights $\widehat{\beta_j^C}$, for reasons outlined in the main text, we recommend the resampling-based (bootstrap) hypothesis test proposed there and in Section 4.3 of this Appendix. In particular, asymptotics may not apply in small studies, and the random variable $\widehat{\beta}$ is dependent on the randomness in $\overline{X^T} - \overline{X^C}$. We thus recommend the bootstrap test, which accounts for this dependence, rather than use of the asymptotic variance of $\widehat{\beta_j^C}$, holding the weights fixed. However, the derivations this section are useful for defining and further highlighting several theoretical features of the statistics.

We also show in subsection 4.2.4 that when covariates are sufficient, we can validly use a test based on δ_{PWLR} to assess as-if random. This is an empirical corollary to Theorem 1 in Section 3.

[11] By the Frisch–Waugh–Lovell (FWL) theorem or "regression anatomy" (Angrist and Pischke 2009: 3.1.2), each element β_j of $p \times 1$ vector of coefficients β in the analogous finite-population regression can be represented as the coefficient from the bivariate regression of $Y(0)$ on the residual of X_j on the other $p - 1$ covariates.

4.3.1 Distribution of the unweighted sum δ_{UW}

Consider as a benchmark the unweighted sum,

$$\delta_{UW} = \sum_{j=1}^{p} \delta_j, \tag{7}$$

where each δ_j is the difference of means across the treatment and control groups on covariate j. Here, "UW" stands for "unweighted." Note that each difference of means δ_j is a random variable, with the randomness induced solely by treatment assignment, and thus so is the sum δ_{UW}.[12] Here, each covariate difference—unlike in our preferred approach—receives the same weight.

The distribution of the random variable δ_{UW} is then as follows. Let $\sigma^2_{X_j}$ denote the variance of the covariate X_j calculated over all N units in the finite population and σ_{X_j,X_k} be the finite-population covariance between covariate X_j and covariate X_k.[13] The sizes of the treatment and control groups are fixed at n_1 and n_0, respectively, with $n_1 + n_0 = N$. For ease of exposition but without loss of generality, suppose $n_1 = n_0$.

Theorem A.1 *(Distribution of the unweighted sum of covariate differences of means). When treatment assignment is randomized, (1) $E(\delta_{UW}) = 0$, and (2) the sum has an exact and fully observable variance*

$$\text{Var}(\delta_{UW}) = \frac{N^2}{N-1}\frac{1}{n_0(n_1)}\left[\sum_{j=1}^{p}\sigma^2_{X_j} + 2\sum_{j<k}^{p}\sigma_{X_j,X_k},\right].$$

Also, (3) δ_{UW} is asymptotically normal.

Before turning to the proof, we note that the variance in Theorem A.1 is exact and fully observable—not estimated from sample data—because we can observe covariate values for every unit in the finite population. Thus, the variances and covariances $\sigma^2_{X_j}$ and σ_{X_j,X_k} can be calculated exactly for all j and all k in a given data set. The variance of the difference of means for each covariate is given by a formula that reflects both sampling without replacement from the finite population and the dependence between the treatment and control group means, similar to the variance of an estimator of an average treatment effect.[14] However, unlike in that case, here there are no unobservable sample covariances because X_i is invariant to treatment assignment.[15] Note also that when each covariate is standardized, so that the finite-population variance of each covariate is equal to 1, we have $\sum_{j=1}^{p}\sigma^2_{X_j} = p$ and each covariance σ_{X_j,X_k} is the coefficient of correlation r between X_j and X_k.

We consider in turn the three claims in the theorem—i.e., regarding (1) the expectation, (2) the variance, and (3) asymptotic normality of the random variable δ_{UW}.

Proof. (1). Each random variable δ_j can be written $(1/n_1)Z'X_j + (1/n_0)(1-Z)'X_j = X_j^T - X_j^C$, where X_j^T

[12] To make the dependence on Z explicit, for example, each random variable δ_j could be written $(1/n_1)Z'X_j + (1/n_0)(1-Z)'X_j = X_j^T - X_j^C$, where X_j^T and X_j^C are the means of covariate j in the treatment and control groups, respectively.

[13] That is, $\sigma^2_{X_j} = \frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X_j})^2$, and $\sigma_{X_j,X_k} = \frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})$.

[14] See Neyman (1923); Freedman (2007: A32-A34); Samii and Aronow (2011, Theorem 2); Gerber and Green (2012: 57); Dunning (2012: 193).

[15] Thus, this is similar to the variance of $\widehat{ATE}$ under the strict null hypothesis that $Y_i(1) = Y_i(0)$ for all i.

and X_j^C are the means of covariate j in the treatment and control groups, respectively. Random assignment of the treatment implies that $Z_i \perp\!\!\!\perp u_i$ for any fixed variate u_i, including each of the X_js. Viewed differently, the treatment and control groups are both simple random samples from the same underlying population. The expectations of the two sample means therefore coincide: $\mathrm{E}(\delta_j) = \mathrm{E}(X_j^T) - \mathrm{E}(X_j^C) = 0$ for $j = 1, \ldots, p$. Thus, after distributing expectations, $\mathrm{E}(\delta_{UW}) = \mathrm{E}(\delta_1) + \mathrm{E}(\delta_2) + \ldots \mathrm{E}(\delta_p) = 0$.

(2). Next, for the variance, consider as a preliminary two arbitrary features (u_i, v_i) in the finite population $i = 1, \ldots, N$. Define the population variances as

$$\sigma_u^2 = \frac{1}{N} \sum_i^N (u_i - \overline{u})^2 \tag{8}$$

and

$$\sigma_v^2 = \frac{1}{N} \sum_i^N (v_i - \overline{v})^2, \tag{9}$$

where $\overline{u} = 1/N \sum_{i=1}^N u_i$ and $\overline{v} = 1/N \sum_{i=1}^N v_i$ are the population means. The population covariance between these features is

$$\sigma_{u,v} = \frac{1}{N} \sum_i^N (u_i - \overline{u})(v_i - \overline{v}). \tag{10}$$

Let $\overline{U}_z$ denote the sample average in the treatment ($z = 1$) or control ($z = 0$) group, and similarly for $\overline{V}_z$.[16] Here, $\overline{U}_z$ and $\overline{V}_z$ are random variables, due to randomness in Z. If we observe both u_i and v_i in the treatment sample, then we have

$$\mathrm{Cov}(\overline{U}_1, \overline{V}_1) = \frac{N - n_1}{N - 1} \frac{\sigma_{u,v}}{n_1} = \frac{n_0}{N - 1} \frac{\sigma_{u,v}}{n_1}, \tag{11}$$

since the features are drawn without replacement from a finite population of size N (Cochran 1977, Theorem 2.3). If we observe u_i and v_i in the control sample, then

$$\mathrm{Cov}(\overline{U}_0, \overline{V}_0) = \frac{N - n_0}{N - 1} \frac{\sigma_{u,v}}{n_1} = \frac{n_1}{N - 1} \frac{\sigma_{u,v}}{n_1} \tag{12}$$

(If $n_1 \neq n_0$, these theoretical covariances must be figured separately for the two groups).

If we observe u_i only for i in the treatment sample and v_i only for i in the control sample, the variances of the samples averages are

$$\mathrm{Var}(\overline{U}_1) = \frac{N - n_1}{N - 1} \frac{\sigma_u^2}{n_1} = \frac{n_0}{N - 1} \frac{\sigma_u^2}{n_1} \tag{13}$$

and

$$\mathrm{Var}(\overline{V}_0) = \frac{N - n_0}{N - 1} \frac{\sigma_v^2}{n_0} = \frac{n_1}{N - 1} \frac{\sigma_v^2}{n_0}. \tag{14}$$

Using combinatorial calculations (see Freedman et al. 2007: A32-34 or Neyman et al. 1923, also Lin et al.

[16] Thus, $\overline{U_1}$ can be written as $\frac{1}{n_1} Z'u$ and $\overline{U_0}$ as $\frac{1}{n_0}(1 - Z)'v$, where u is an $N \times 1$ vector collecting the N values of u_i; we can write $\overline{V_z}$ similarly. This notation clarifies that randomness in the sample means depends on treatment assignment Z.

2023), the covariance of the sample averages is

$$\text{Cov}(\overline{U}_1, \overline{V}_0) = -\frac{\sigma_{u,v}}{N-1}. \tag{15}$$

The variance of the difference of the sample means $\overline{U}_1 - \overline{V}_0$ is then

$$\begin{aligned}
\text{Var}(\overline{U}_1 - \overline{V}_0) &= \text{Var}(\overline{U}_1) + \text{Var}(\overline{V}_0) - 2\text{Cov}(\overline{U}_1, \overline{V}_0) \\
&= \frac{1}{N-1}\left[\frac{n_0\,\sigma_u^2}{n_1} + \frac{n_1\,\sigma_v^2}{n_0} + 2\,\sigma_{u,v}\right] \\
&= \frac{1}{N-1}\left[\frac{n_0^2\,\sigma_u^2 + n_1^2\,\sigma_v^2 + 2n_1(n_0)\,\sigma_{u,v}}{n_1(n_0)}\right],
\end{aligned} \tag{16}$$

using (13), (14), and (15) in the second step. (For related derivations, see Neyman et al. 1923; Freedman et al. 2007: A32-A34); Samii and Aronow 2012: Theorem 2; Gerber and Green 2012: 57; or Dunning 2012: 193.[17]

With these preliminaries, we can derive the variance of the random variable δ_{UW}. We have

$$\begin{aligned}
\text{Var}(\delta_{UW}) &= \text{Var}(\sum_{j=1}^{p} \delta_j) \\
&= \sum_{j=1}^{p} \text{Var}(\delta_j) + 2\sum_{j<k}^{p} \text{Cov}(\delta_j, \delta_k).
\end{aligned} \tag{17}$$

First, $\text{Var}(\delta_j)$ has the same form as the variance of $\overline{U_1} - \overline{V_0}$ when $u_i = v_i$ for all i (since X_{ji} has the same value whether unit i is assigned to the treatment or the control group). Using equation (16), we find

$$\begin{aligned}
\text{Var}(\delta_j) &= \text{Var}(\overline{X}_{j1} - \overline{X}_{j0}) \\
&= \frac{1}{N-1}\left[\frac{n_1\,\sigma_{X_j}^2}{n_0} + \frac{(n_0)\,\sigma_{X_j}^2}{n_1} + 2\,\sigma_{X_j}^2\right] \\
&= \frac{1}{N-1}\left[\frac{n_1^2\,\sigma_{X_j}^2 + n_0^2\sigma_{X_j}^2 + 2n_0(n_1)\,\sigma_{X_j}^2}{n_0(n_1)}\right] \\
&= \frac{1}{N-1}\left[\frac{(n_0+n_1)^2\sigma_{X_j}^2}{n_0(n_1)}\right]. \\
&= \frac{1}{N-1}\left[\frac{N^2\sigma_{X_j}^2}{n_0(n_1)}\right].
\end{aligned} \tag{18}$$

Here, $\overline{X}_{j1}$ indicates the sample average of X_j in the treatment group and $\overline{X}_{j0}$ indicates the sample average in the control group.[18] Also, $\sigma_{X_j}^2$ is (8) with $u_i = X_{ij}$: it denotes the variance of the covariate X_j calculated

[17] Following the previous note, the difference of means $\overline{U_1} - \overline{V_0}$ can be written as $\delta_{u,v} = \frac{1}{n_1}Z'u + \frac{1}{n_0}(1-Z)'v$, where u and v are the $N \times 1$ vectors collecting the N values of u_i and v_i, respectively.

[18] We could write $\overline{X}_{j1} = \frac{1}{n_1}Z'X_j$ and $\overline{X}_{j0} = \frac{1}{n_0}(1-Z)'X_j$ to clarify dependence of the sample averages on the random assignment

over all N units in the finite population, that is,

$$\sigma^2_{X_j} = \frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X}_j)^2, \tag{19}$$

where $\overline{X}_j$ is the mean of X_j over the N study units. Also, since the covariance of a variable with itself is its variance, $\text{Cov}(X_j, X_j) = \sigma^2_{X_j}$.

Thus, we can calculate an exact, fully observable sampling variance for each δ_j. As under a strict null hypothesis, where one "sees" potential outcomes for unit i in both treatment and control conditions (by the stipulation that $Y_i(1) = Y_i(0)$ for all i), here we observe covariate values X_i under both treatment and control conditions, whether unit i is in fact assigned to the treatment or the control group—because covariates are fixed values invariant to treatment assignment.[19] Note also that σ^2_k is fully observed because we see values of each covariate for every study unit. In sum, there are no terms in (18) or (27) that would need to be estimated from sample data: this exact variance is fully observable.

As for $\text{Cov}(\delta_j, \delta_k)$, we have

$$\begin{aligned}\text{Cov}(\delta_j, \delta_k) &= \text{Cov}(\overline{X}_{j1} - \overline{X}_{j0}, \overline{X}_{k1} - \overline{X}_{k0}) \\ &= \text{Cov}(\overline{X}_{j1}, \overline{X}_{k1}) - \text{Cov}(\overline{X}_{j1}, \overline{X}_{k0}) - \text{Cov}(\overline{X}_{j0}, \overline{X}_{k1}) + \text{Cov}(\overline{X}_{j0}, \overline{X}_{k0}).\end{aligned} \tag{20}$$

The first and fourth terms in (20) are the covariances of the sample averages of two features, both sampled without replacement from a finite population of size N. Using (11) and (12), we have

$$\text{Cov}(\overline{X}_{j1}, \overline{X}_{k1}) = \frac{N - n_1}{N - 1}\frac{\sigma_{X_j,X_k}}{n_1} = \frac{n_0}{N - 1}\frac{\sigma_{X_j,X_k}}{n_1} \tag{21}$$

and

$$\text{Cov}(\overline{X}_{j0}, \overline{X}_{k0}) = \frac{N - n_0}{N - 1}\frac{\sigma_{X_j,X_k}}{n_0} = \frac{n_1}{N - 1}\frac{\sigma_{X_j,X_k}}{n_0}. \tag{22}$$

The second and third terms in (20) are instead the covariances of the sample averages of two features, one assigned to the treatment group and one assigned to the control group. Using (15), we have

$$\text{Cov}(\overline{X}_{j1}, \overline{X}_{k0}) = \text{Cov}(\overline{X}_{j0}, \overline{X}_{k1}) = -\frac{1}{N - 1}\sigma_{X_j,X_k}, \tag{23}$$

where σ_{X_j,X_k} is the population covariance given in (10), with the covariates X_j and X_k in place of u and v.

vector Z.

[19] If $u_i = Y_i(1)$ is a potential outcome under treatment and $v_i = Y_i(0)$ is a potential outcome under control, then the random variable $\delta_{u,v}$ estimates the average treatment effect. Then $\text{Var}(\delta_{u,v})$ is the variance of $\widehat{ATE}$ under a strict null hypothesis of no unit-level effect.

Thus,

$$\begin{aligned}
\mathrm{Cov}(\delta_j, \delta_k) &= \frac{n_0}{N-1}\frac{\sigma_{X_j,X_k}}{n_1} + 2\frac{\sigma_{X_j,X_k}}{N-1} + \frac{n_0}{N-1}\frac{\sigma_{X_j,X_k}}{n_0} \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0}{n_1} + 2 + \frac{n_1}{n_0}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0^2}{n_1(n_0)} + \frac{2n_1(n_0)}{n_1(n_0)} + \frac{n_1^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0^2 + 2n_1(n_0) + n_1^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{(n_0+n_1)^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}].
\end{aligned} \tag{24}$$

Returning to (17) and substituting for $\mathrm{Var}(\delta_j)$ and $\mathrm{Cov}(\delta_j, \delta_k)$, we have

$$\begin{aligned}
\mathrm{Var}(\delta_{UW}) &= \mathrm{Var}(\sum_{j=1}^{p}\delta_j) \\
&= [\sum_{j=1}^{p}\mathrm{Var}(\delta_j) + 2\sum_{j<k}^{p}\mathrm{Cov}(\delta_j, \delta_k)] \\
&= [\sum_{j=1}^{p}\frac{1}{N-1}\left[\frac{N^2\sigma_{X_j}^2}{n_0(n_1)}\right] + 2\sum_{j<k}^{p}\frac{\sigma_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}] \\
&= [\frac{N^2}{N-1}\left[\frac{1}{n_0(n_1)}\sum_{j=1}^{p}\sigma_{X_j}^2\right] + 2\frac{N^2}{N-1}\frac{1}{n_1(n_0)}\sum_{j<k}^{p}\sigma_{X_j,X_k}] \\
&= \frac{N^2}{N-1}\frac{1}{n_0(n_1)}\left[\sum_{j=1}^{p}\sigma_{X_j}^2 + 2\sum_{j<k}^{p}\sigma_{X_j,X_k}\right].
\end{aligned} \tag{25}$$

Thus, data on p covariates for N units allows us to calculate the exact variance of the sum of the covariate differences of means. As with the variance of each δ_j, $\mathrm{Var}(\delta)$ is fully observable: it need not be estimated from sample data because $\sigma_{X_j}^2$ and $\sigma_{k,j}$ are both measurable from the covariate data for the N units in the population. Note also that here we assume that n_1 and n_0 are fixed, not random.[20]

Note that when the covariate X_j is standardized as

$$(X_{ij} - \overline{X_j})/\sigma_j, \tag{26}$$

[20]This is standard in experimental analysis, where the group sizes are planned in advance of randomization; for a natural experiment, the assumption is more debatable. If the group sizes are random variables, ratio-estimator bias may arise for small samples, though with moderately large n_1 and n_0 the distinction should make little difference.

we find using (18) that

$$\text{Var}(\delta_{j,\text{stand}}) = \frac{N^2}{N-1}\left[\frac{1}{n_0(n_1)}\right], \tag{27}$$

and σ_{X_j,X_k} in (24) is ρ_{X_j,X_k}, the correlation of $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$. Then

$$\text{Var}(\delta)_{UW,\text{stand}}) = \frac{N^2}{N-1}\frac{1}{n_0(n_1)}\left[p + 2\sum_{j<k}^{p}\rho_{X_j,X_k}\right]. \tag{28}$$

(3). Finally, for the third claim in the theorem, note that under an appropriate central limit theorem (Erdős and Rényi 1959, Hájek 1960, Höglund 1978), the sampling distribution of each δ_j and thus of δ is asymptotically normal. It will be approximately normal in a finite study group if n_0 and n_1 are large or even moderately sized, and even more so if the variables X_j themselves have an approximately normal distribution. That each δ_j is a difference of averages also helps foster approximate normality, even in small samples. In sum, $\delta \dot{\sim} \quad N(0, \text{Var}(\delta))$, where here $\dot{\sim}$ means "approximately distributed as," which can aid hypothesis testing when justified.

4.3.2 Hotelling's T^2 statistic, δ_{UW}, and the F-distribution

The distribution of δ_{UW}, the unweighted sum of covariate differences of means, is closely related to Hotelling's (1931) two-sample T^2 statistic; indeed, the latter simply normalizes the former by the inverse of the pooled sample covariance matrices, producing possible efficiency gains, as we show in this subsection. Hotelling's T^2 can in turn readily be related to the F-distribution used in some multivariate (unweighted) covariate balance tests.

First, define $\delta^2_{UW} = \sum_{j=1}^{p}\delta_j^2$, where as in the text each δ_j is $\overline{X_j^T} - \overline{X_j^C}$, i.e. the difference of means on covariate j. In vector notation, this can be written as

$$\delta^2_{UW} = (\overline{X^T} - \overline{X^C})'(\overline{X^T} - \overline{X^C}),$$

where e.g. $\overline{X^T}$ is the $p \times 1$ vector of means of the p covariates in the treatment group. As random variables, the covariate means (and their difference) may be approximately normally distributed in finite samples, and they are asymptotically normal by an appropriate central limit theorem (see point (3) in the proof of Theorem A.1). Thus, the sum of squared differences, δ^2_{UW}, is approximately χ^2.

Hotelling's T^2, by contrast, can be written in our context as

$$t^2 = \frac{n_0 n_1}{N}(\overline{X^T} - \overline{X^C})'\widehat{\Sigma}^{-1}(\overline{X^T} - \overline{X^C}),$$

where $\widehat{\Sigma}$ is the $p \times p$ pooling sample variance-covariance matrix. Note that pooling across the treatment and control groups makes sense here: $\boldsymbol{X_i}$ is the same whether i is assigned to the treatment or control group, so the sample means for each group are drawn from the same distribution. Thus, t^2 is essentially δ^2_{UW}, scaled by a constant of proportionality and the inverse of the pooled variance-covariance matrix, which may make it more efficient. Hotelling's statistic is distributed as a T^2 random variable with parameter p and $N - 2$ degrees of freedom.

Finally,

$$\frac{(N-p-1)}{(N-2)p}t^2 \sim F_{(p,N-p-1;t^2)}.$$

Unweighted statistics such as δ_{UW} or t^2, like an F-test after regression of treatment assignment on covariates, may perform somewhat differently, as our simulations in the paper suggest, but they are all unresponsive to the level or distribution of covariate prognosis.

4.3.3 Conditional distribution of δ_{PWLR}

We can now relate the conditional distribution of δ_{PWLR}—also a random variable, with randomness due to treatment assignment—to the unweighted sum:

Theorem A.2 *(Distribution of the prognosis-weighted sum of covariate differences of means, δ_{PWLR}). When treatment is randomly assigned, (1)* $\text{plim}(\delta_{\text{PWLR}}) = 0$*. Also, (2) the large-sample variance of δ_{PWLR}, conditional on the weights, is proportional to* $\text{Var}(\delta_{UW})$ *as given in Theorem A.1.*

Proof. Consider the distribution of

$$\delta_{PWLR} = (\overline{X^T} - \overline{X^C})'\widehat{\beta^C},$$

that is,

$$\widehat{\overline{Y(0)^T}} - \overline{Y(0)^C},$$

the difference between the fitted average potential outcomes under control in the treatment and control groups.

(1). By Slutksy's theorem, it is immediate that $\text{plim}(\delta_{PWLR}) = \text{plim}(\overline{X^T} - \overline{X^C})'\widehat{\beta^C} = \text{plim}(\overline{X^T} - \overline{X^C})'\text{plim}(\widehat{\beta^C}) = 0$ when treatment is randomly assigned; in that case, the covariate means in the treatment and control groups are equal in expectation.

(2). As for the conditional variance of δ_{PWLR},

$$\begin{aligned}
\text{Var}(\delta_{PWLR}|\widehat{\beta}) &= \text{Var}(\sum_{j=1}^{p}\widehat{\beta_j^C}\delta_j|\widehat{\beta}) \\
&= \sum_{j=1}^{p}\text{Var}(\widehat{\beta_j^C}\delta_j|\widehat{\beta}) + 2\sum_{j<k}^{p}\text{Cov}(\widehat{\beta_j^C}\delta_j, \widehat{\beta_k}^C\delta_k|\widehat{\beta}) \\
&= \sum_{j=1}^{p}\widehat{\beta_j^C}^2\text{Var}(\delta_j) + 2\sum_{j<k}^{p}\widehat{\beta_j^C}\widehat{\beta_k^C}\text{Cov}(\delta_j, \delta_k) \\
&= \sum_{j=1}^{p}\widehat{\beta_j^C}^2\frac{1}{N-1}\left[\frac{N^2\sigma^2_{X_j}}{n_0(n_1)}\right] + 2\sum_{j<k}^{p}\widehat{\beta_j^C}\widehat{\beta_k^C}\frac{\sigma_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}],
\end{aligned}$$

where in the final line we use (17) and (24). When the elements of $\boldsymbol{X}$ are standardized, we have

$$\text{Var}(\delta_{RW,\text{stand}}|\widehat{\beta}) = \sum_{j=1}^{p}\widehat{\beta_j^C}^2\frac{1}{N-1}\left[\frac{N^2}{n_0(n_1)}\right] + 2\sum_{j<k}^{p}\widehat{\beta_j^C}\widehat{\beta_k^C}\frac{\rho_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}] \tag{29}$$

In sum, the conditional variance of δ_{PWLR} is proportional to the variance of δ_{UW}, with constants of proportionality equal to the regression weights. With standardized regressions, terms for which the fitted regression coefficients approach zero will vanish.

Theorem A.2 gives a large-sample result on the conditional distribution of our test statistic. To conduct hypothesis tests, we could form a t- or z-test based on the ratio of δ_{PWLR} to the square root of an estimator of the conditional variance derived in the proof of theorem, e.g., equation (29). However, such a large-sample approximation may not be reliable in small studies.[21] Moreover, the conditional distribution does not readily account for randomness in the weights (that is, the regression coefficients fit in the randomly assigned control group). We therefore instead recommend the bootstrap hypothesis test we develop in section 4.4.

4.3.4 Testing as-if random with δ_{PWLR} and sufficient covariates

When $\boldsymbol{X}$ is sufficient, the non-independence of treatment assignment and a consistent estimator of the conditional expectation of potential outcomes implies the non-independence of treatment assignment and potential outcomes. In other words, we can use a consistent estimator of the finite-population regression to test as-if random.

Let $Y_i(0)_{lr} = X_i\beta$ be the finite-population regression (with "lr" for linear regression), and $\widehat{Y_i(0)_{lr}} = X_i^{C\prime}\widehat{\beta^C}$ be the regression of outcomes on covariates in the control group, i.e. the sample version of this regression, where "C" stands for the control group.

Then we have the following theorem.

Theorem A.2 *(Observable Implication of As-If Randomization). Suppose that $\boldsymbol{X}$ is sufficient for $Y(0)$, $Y_i(0)_{lr} = X_i\beta$, and $\widehat{Y(0)_{lr}}$ is a consistent estimator for $Y_i(0)_{lr}$. Then: $Z \not\perp\!\!\!\perp \widehat{Y(0)_{lr}} \implies Z \not\perp\!\!\!\perp Y(0)$.*

Proof. Consider the estimator $\widehat{Y(Z)|X} = \sum_j \widehat{\beta_j^Z} X_j$, which is the sample analogue of $Y(Z)_{lr}$, as defined in equation (2) in the text for $Y(0)$; randomness in $\widehat{\beta_j^Z}$ is induced by treatment assignment. Define, for any $U \neq \boldsymbol{X}$, a linear estimator $\widehat{Y(Z)|X, U}$ such as $\widehat{\beta_j^Z} X_j + \widehat{\gamma} U$.

By sufficiency, $\forall U \neq X, Y(Z) \perp\!\!\!\perp U|\boldsymbol{X}$. Also, by the Frisch-Waugh-Lovell theorem (or "regression anatomy," Angrist and Pischke 2009: 3.1.2), the coefficient on U in the sample regression is

$$\widehat{\gamma} = \frac{\widehat{\text{Cov}}(\boldsymbol{Y(0)}, \widehat{\widetilde{\boldsymbol{U}}})}{\widehat{\text{Var}}(\widehat{\widetilde{\boldsymbol{U}}})}, \tag{30}$$

where $\widehat{\widetilde{\boldsymbol{U}}}$ is the residual from the sample regression of U on $\boldsymbol{X}$ and we use $\widehat{\ }$ to denote the sample estimator.

By consistency of $\widehat{Y(Z)}$ (from Theorem A.2) and sufficiency, $\text{E}(\widehat{\gamma}) \to 0$, and

$$\lim_{n\to\infty} P\left(\left\{|\widehat{Y(Z)|X} - \widehat{Y(Z)|X, U}| > 0\right\} > \epsilon\right) \to 0$$

[21] Inter alia, there may be ratio-estimator bias (the sample regression estimator can be viewed as a ratio of random variables, since covariate values in the control group are random).

for arbitrarily small ϵ. Take U to be X^C, the complement of $\boldsymbol{X}$. Then we have

$$\begin{aligned} Z \not\perp \widehat{Y(Z)|X} &\implies \\ Z \not\perp \widehat{Y(Z)}|X, X^C &\implies \\ Z \not\perp \widehat{Y(Z)} &\implies \\ Z \not\perp Y(Z). & \end{aligned}$$

The theorem provides an empirical corollary to Theorem 1: when $\boldsymbol{X}$ is sufficient, we can validly use the empirical prognosis-weighted statistic to test as-if random.

Note also that δ_{PWLR} is just $\widehat{Y(0)_{lr}}$ in the treatment group minus $\widehat{Y(0)_{lr}}$ in the control group, and both terms are consistent for $X_i\beta$ under as-if random: see the proof of claim (1) in Theorem A.1.

4.4 A bootstrapped prognosis-weighted test of as-if random

For hypothesis testing, we propose a resampling (a.k.a. bootstrap) technique which allows comparison of the observed value of a test statistic to its exact randomization distribution.[22]

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. Thus, for δ_{PWLR}, we draw two independent samples of potential outcomes from the control group; fit the prognosis regression in one of them; calculate a bootstrap test statistic, i.e., the prognosis-weighted difference of means; and repeat the bootstrap B times in order to compare an observed test statistic to its randomization distribution. For non-linear fitting methods, the procedure is parallel but uses the chosen non-linear regression or machine learning procedure in place of linear regression.

The validity of this procedure rests on two key features. First, the expectation of the covariate difference of means e.g. in δ_{PWLR} is zero, as it is when treatment assignment is randomized. Thus, we compare the expected values of averages of two independent samples drawn from the same finite bootstrap population.[23] Second, the procedure allows in a natural way for the statistical dependence between the random variable $\widehat{\beta^C}$—as realized in the control group—and $\overline{X^C}$, with treatment assignment as the only source of stochastic variation. Note that the bootstrap uses only values of $Y(0)$ from the control group to simulate the distribution of prognosis weights.

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).

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, $\overline{X^{C*}}$.

[22] On randomization tests, see Fisher (1935); also inter alia Caughey et al. (2017).

[23] 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 et al. 2007: A32-A34; Samii and Aronow 2012, Theorem 2; Gerber and Green 2012: 57; or Dunning 2012: 193).

2. Take another independent sample with replacement, *also from the observed control group*, and calculate the sample mean of the covariates, $\overline{X^{T*}}$.

3. Calculate a simulated $\delta_{PW}^{*b} = (\overline{X^{T*}} - \overline{X^{C*}})'\widehat{\beta^{C*}}$.

4. Repeat steps (1)-(3) B times ($B = 500$ in our default).

5. Calculate a two-sided randomization-based p-value as

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\delta_{PWLR}^{*b}| \geq |\delta_{PWLR}^{\text{obs}}|), \tag{31}$$

where $\mathbb{I}$ is an indicator function that equals 1 if its argument is true and 0 otherwise, and $\delta_{PWLR}^{\text{obs}}$ is the observed value of the prognosis-weighted test statistic. Reject the null if, e.g., $p^* < 0.05$.

4.5 Machine-learning methods for fitting δ_{PW}

The fitted value approach also leads naturally to alternative, more flexible nonlinear techniques. The predicted potential outcomes in δ_{PW} in equation (3) can be formed by a host of methods.

In subsection **??** and Online Appendix Section 7, we explore the performance of two main alternatives. First, we extend the linear regression-based approach of subsection **??** to include polynomial terms and a full set of covariate interactions. Thus, a fully "saturated" regression produces the fitted values.

Second, we extend our software `pwtest` to allow for a host of more flexible methods, including machine learning (ML) techniques. The options include, among others, generalized linear models with LASSO, Bayesian Additive Regression Trees (BART), random forests, and gradient boosted trees. The strategy is the same across all methods and follows the following steps:

1. Fit $\widehat{Y(0)^C}$ on covariate set X^C (i.e., subsetting to control units), using a given method;

2. With the resulting fit, obtain $\widehat{Y^T(0)}$ using treatement-group covariate values X^T; and

3. Calculate the observed δ_{PW} as defined by equation (3).

The software bootstraps a hypothesis test and associated p-values using the approach described in subsection 4.4 and returns diagnostic measures of prognosis.

Our observed δ_{PW} statistic is defined as

$$\delta_{PW} = \overline{\widehat{Y(0)^T}} - \overline{Y(0)^C} \tag{32}$$

or the difference between the fitted average potential outcomes under control in the treatment and control groups, where fitted values of potential outcomes under control for treatment units can be obtained by fitting a procedure $\mathscr{M}$ on the units assigned to treatment, such that

$$\widehat{Y(0)^T} = \mathscr{M}(X^T)$$

Under the linear regression approach we have considered in our test of as-if random described in Section 4.3.3, where the test statistic is δ_{PWLR}, we define $\mathscr{M}(X^T) = X^T\widehat{\beta^C}$. In this section, we describe our approach under alternative definitions of $\mathscr{M}$.

Our approach can draw from a variety of more flexible methods for fitting $\widehat{Y^T}$, including generalized linear models with LASSO, random forests, Bayesian Additive Regression Trees (BART), gradient boosting frameworks, and others.

The standard estimation strategy for δ_{PW} is the same across all methods and follows the following steps:

Step 1 Train a model $\mathscr{M}$ to fit $Y^C(0)$ on covariate set X^C. This model fitting step subsets to observed $Y(0)$ (i.e., control units). In machine learning methods, this subset of the data will consist of the training set.

Step 2 With the resulting model, obtain $\widehat{Y^T(0)}$ for the subset of the data assigned to treatment using observed covariate values X^T (i.e. the test set).

Step 3 Calculate the observed δ_{PW} as defined by equation (32).

In subsection **??**, we use simulations to assess the performance of the saturated regression and two widely used ML methods—gradient boosted trees and random forests (**????**)—as well as the performance of a procedure for choosing the "best"-fitting model that we discuss next.

4.5.1 Hypothesis testing with δ_{PWML}

Step 4 Draw a sample with replacement from the observed control group. Fit $\mathscr{M}$ on this sample to obtain a $\widehat{Y^{C*}(0)}$.

Step 5 Take another independent sample with replacement, *also from the observed control group*. Fit $\mathscr{M}$ on this bootstrap sample to obtain $\widehat{Y^{T*}(0)}$.

Step 6 Calculate $\delta^*_{PW} = \widehat{\overline{Y^{T*}(0)}} - \widehat{\overline{Y^{C*}(0)}}$

Step 7 Repeat Steps (4)-(6) B times ($B = 500$ in our default).

Step 8 Calculate a two-sided randomization-based p-value as

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\delta^{*b}_{PW}| \geq |\delta^{\text{obs}}_{PW}|), \tag{33}$$

where $\mathbb{I}$ is an indicator function that equals 1 if its argument is true and 0 otherwise, and δ^{obs}_{PW} is the observed value of the prognosis-weighted test statistic. Reject the null if, e.g., $p^* < 0.05$.

4.5.2 Cross-validation and choice of methods

Our `pwtest` function also allows for an automated selection of the method with the best predictive performance. In this case, the method that predicts $Y(0)$ most accurately from covariates in the control (or training) group is selected for use in the resulting test procedure. For the ML methods, this is also based on a k-fold cross-validation process for selection of hyperparameters using control group units only. To select an appropriate fitting procedure in a data-driven way, the software picks the estimation method with

the highest R^2 on the task $Y(0)|X$ in the control group. We discuss further details in subsections 4.4 and 7.5 of the Appendix.

Ideally, the procedure for selecting the fitting procedure should be pre-specified in advance of testing. However, we also note that for reasons described in subsection **??**—especially simplicity and interpretability of the weights—there may often be a rationale for using the baseline linear approach (the default in `pwtest`), even if non-linear methods can provide a slight improvement in power, and we recommend also reporting tests using this simple approach. We return to discussion of this point in connection with the simulation results in subsection **??**.

5 Prognosis-weighted tests in regression-discontinuity designs

In this section, we propose tests for two identification conditions in RD designs: as-if random and continuity of potential outcomes.

The approach for continuity of potential outcomes, as discussed in section 4.2 of the paper, involves testing for equality of the intercepts of two prognosis-weighted regressions, one on each side of the assignment threshold. Here, we give more details on that procedure and describe two approaches to statistical inference and hypothesis testing.

As we also discuss, there are settings in which it makes sense to test the stronger assumption of as-if random; we discuss prognosis-weighted tests of as-if random in subsection 5.2. The approach for as-if random is similar to that developed in section 4.1 in the paper, including the bootstrap in section 4.1.1. However, it requires specifying a procedure for bandwidth selection (i.e., the size of the window around the key RD threshold that will define the study group for testing).

5.1 Testing continuity of potential outcomes in RD designs

Analysts have rightly noted that in many RD designs, as-if random should be replaced with the (weaker) assumption that the regression functions relating potential outcomes to the forcing variable (a.k.a. the running variable or "score") are continuous at the threshold determining treatment assignment (Calonico et al. 2014; De la Cuesta and Imai 2016). This may especially be so when the slope of the regression function relating potential outcomes to the forcing variable is not flat (see Dunning 2012 Chapters 3 and 5; Cattaneo et al. 2015; Sekhon and Titiunik 2017). When the forcing variable has a strong relationship to the response variable, we might expect as-if random to fail, i.e., we may expect to find average differences in potential outcome even in narrow bandwidths around the threshold determining treatment assignment. However, functions of average potential outcomes may nonetheless be continuous at the threshold.

Our example of close-election designs discussed in section 5.1 of the text provides one illustration. At least in the U.S. House context, party vote share at time t (the running variable in many studies) appears strongly related to party vote share or incumbency at time $t + 1$ (the outcome). (Per our discussion of the weak prognosis of lagged incumbency generally, this is not the case in all election contexts). Thus, because the slope of the potential outcome regression function is increasing, we may expect average differences of potential outcomes in narrow bandwidths on each side of the threshold. However, those functions may be continuous at the threshold. In the U.S. House, there are substantial differences in the average values of covariates related to potential outcomes in narrow bandwidths around the assignment threshold, yet, consistent with De la Cuesta and Imai (2016), our evidence from prognosis-weighted tests is consistent with continuity of potential outcomes at the threshold.

In this case, the key condition to test is not as-if random (Assumption 1) but rather:

Assumption 2. *(Continuity of Potential Outcomes—RD Designs) Potential outcomes regression functions are continuous at the threshold determining treatment assignment.*

Continuity implies that the limits of the regression functions are the same approaching from above and below the threshold. This motivates the standard approach of testing for the equality of intercepts of two regressions, fit above and below the threshold value of the running covariate.

However, researchers typically test for the continuity not of potential outcomes—but of *covariates*. Thus, they regress each pre-treatment covariate separately on the forcing variable, above and below the RD threshold, and conduct a test for equality of the intercepts at the threshold.

Unfortunately, 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 (section 3.2). Covariates may be continuous at the threshold and yet potential outcomes may not be; or vice versa. The standard approach also raises the problems of indeterminacy and multiple testing (De la Cuesta and Imai 2016), as in covariate-by-covariate tests of as-if random.

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, we first project the outcome variable on covariates on the control group side of the RD threshold. Then, we fit regressions—not of covariates, as in standard practice, but of fitted potential outcomes—on the running variable, on each side of the threshold.

In the next sub-section, we provide more details on construction of the test statistic described in the text, while the following sub-section turns to statistical inference and hypothesis testing.

5.1.1 Test statistic: the prognosis-weighted difference of intercepts

Continuity implies that the limits of these functions approaching the threshold from above and below are the same. For $Y(0)$, for example,

$$\lim_{r \downarrow c} \mathbb{E}[Y_i(0)|R_i = r] = \lim_{r \uparrow c} \mathbb{E}[Y_i(0)|R_i = r]. \tag{34}$$

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.

However, researchers typically test for the continuity not of $Y(0)$—as in equation (34)—but of *covariates*. Thus, they regress each placebo or pre-treatment covariate separately on the forcing variable, above and below the RD threshold, and conduct a test for equality of the intercepts at the threshold. Yet, 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 (section 3.2). Covariates may be continuous at the threshold and yet potential outcomes 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. Now, we fit regressions—not of covariates, as in standard practice, but of fitted potential outcomes—on the running variable R_i. Thus the regressions are as follows. First,

$$(\widehat{\alpha_0},\, \widehat{\beta_0}) = \underset{\alpha_0,\beta_0}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c_0 \leq R_i \leq c\}\{\widehat{Y_i(0)} - \alpha_0 - \beta_0(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{35}$$

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). Here, R_i is the forcing variable, c is its value at the assignment threshold, and c_0 is the value that defines the edge of the control-group bandwidth. Similarly,

$$(\widehat{\alpha_1},\, \widehat{\beta_1}) = \underset{\alpha_1,\beta_1}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c < R_i \leq c_1\}\{\widehat{Y_i(0)} - \alpha_1 - \beta_1(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{36}$$

is the intercept and slope from the regression on the treatment-group side, including units up to c_1.[24] For clarity, in (35) and (36), 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 centered value of the forcing variable, $R_i - c$. Note, however, that the latter are distinct from the fitted coefficients of the regression of $Y(0)$ on *covariates* $\boldsymbol{X}$. These, which we label $\widehat{\beta^C}$ as before, are fit in the prognosis regression in the previous step.

Conceptually, it is as if we regressed each pre-treatment covariate on the forcing variable in windows below and above the RD threshold c, as in standard practice (see subsection **??**. However, we combine the intercepts of these separate regressions into one omnibus prognosis-weighted test statistic,

$$\delta_{PW}^{RD} \equiv \widehat{\alpha_1} - \widehat{\alpha_0}, \tag{37}$$

where $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$ are the intercepts at the assignment threshold of the regressions of $\widehat{Y(0)}$ on the forcing variable, on the control-group and treatment-group sides respectively. We can then test the null hypothesis that the expectation of this difference is zero against the alternative of a non-zero difference, or we can flip the null and alternative, as in equivalence testing.

This test of continuity of potential outcomes—as with the test of as-if random—projects out irrelevant covariates and thus bases assessment on the most informative covariates.

5.1.2 A prognosis-weighted sum of intercepts

As we noted in subsection 5.1.1, to form the prognosis-weighted difference of intercepts, it is conceptually as if

- we regressed each pre-treatment covariate on the forcing variable in windows below and above the RD threshold c, as in standard practice; and then
- combined the separate intercepts from these regressions into a single prognosis-weighted difference of intercepts.

[24] As recommended by Calonico et al. (2014) and Cattaneo et al. (2020), equations (35) and (36) 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`.

Indeed, this is true mathematically, as long as the running variable is scaled so that the running covariate is centered at the assignment threshold c. Note first that solving equation (35) gives the least-squares solution

$$\begin{aligned}\widehat{\alpha_0} &= \widehat{Y_i(0)} - \widehat{\beta_0}(R_i - c) \\ &= X_i\widehat{\beta^C} - \widehat{\beta_0}(R_i - c),\end{aligned}$$

where in the second line we plug in the fitted value $\widehat{Y_i(0)} = X\widehat{\beta^C}$. Taking averages over N, we have

$$\begin{aligned}\frac{\sum \widehat{\alpha_0}}{N} &= \frac{\sum \widehat{X_i\beta^C}}{N} - \frac{\sum \widehat{\beta_0}(R_i - c)}{N} \\ &= \overline{X\widehat{\beta^C}} - \widehat{\beta_0}(\overline{R_i} - c)\end{aligned}$$

Thus, when R_i is centered at the assignment threshold so that $\overline{R_i} = c$,

$$\widehat{\alpha_o} = \overline{X\widehat{\beta^C}}. \tag{38}$$

Now, define the "prognosis-weighted sum of intercepts" as

$$\widehat{\alpha_{PW}} = \sum_{k=1}^{p} \widehat{\alpha_k}\widehat{\beta_k^C}, \tag{39}$$

where each $\widehat{\alpha_k}$ is the intercept from the regression of the kth covariate on the forcing variable R_i. That is, teh prognosis-weighted sum is the sum of the intercepts from each of the separate covariate-by-covariate regressions, weighted by the covariate's coefficient in the prognosis regression. Thus,

$$\widehat{\alpha_k} = X_{i,k} - b_k(R_i - c),$$

where b_k is the coefficient of regression of X_k on $R_i - c$. If R_i is centered at c, then taking averages again over N we have

$$\widehat{\alpha_k} = \overline{X_k}. \tag{40}$$

So then plugging (40) into (39) and using (38), we have

$$\begin{aligned}\widehat{\alpha_{PW}} &= \sum_{k=1}^{p} \overline{X_k}\widehat{\beta_k^C} \\ &= \widehat{\alpha_0}.\end{aligned}$$

A similar argument holds for $\widehat{\alpha_1}$, the solution to equation (36). Thus, the test statistic δ_{PW}^{RD} in (37) is the difference of the prognosis-weighted sum of the intercepts from the treatment and control group regressions.

5.1.3 Further details on prognosis-weighted difference of intercepts

In sum, we adapt our procedure as follows to test for continuity of potential outcomes as follows.

First we fit a prognosis regression—the projection of potential outcomes under control onto covariates—on the control group side of the threshold. We then test the continuity of potential outcomes under control

using fitted values from this regression on both sides of the threshold; in our default approach, we compare the intercepts of kernel-weighted local-linear regressions above and below the threshold. This allows us to form the test statistic δ^{RD}_{PWLR}, i.e., the difference of fitted intercepts in equation (9) in the text.

The test statistic is constructed in the following two steps:

1. **Prognosis regression**. First, estimate the vector β by regressing potential outcomes under control on the matrix of pre-treatment covariates, on the control-group side of the RD threshold. Covariates and the outcome may be standardized before running the prognosis regression. This allows us to form the fitted values $\widehat{Y(0)} = X\widehat{\beta^C}$, where $\widehat{\beta^C}$ is the coefficient from the prognosis regression.

   In our default approach, we use all of the units on the control group side of the threshold (i.e., over the support of the running variable on this side of the threshold) to fit the prognosis regression. It is not clear, especially for covariates that are unrelated to the running variable, that units closer to the RD threshold as defined by their values of the running variable will allow us a better approximation of the finite-population relationship between $Y(0)$ and X (note that this is the goal—we are not concerned here with the relationship between $Y(0)$ and the running covariate R_i). Ceteris paribus, use of all observed $Y(0)$ values also reduces statistical uncertainty in the estimate of β, the finite-population regression coefficient. However, users can alter this option by specifying manually a bandwidth for the prognosis regression using our `R` package. In practice, the estimate $\widehat{\beta^C}$ from the prognosis regression may not be very sensitive to the bandwidth one chooses. It is critical, however, that only units on the control group side of the threshold are used in estimating the finite-population relationship between $Y(0)$ and β—since on the other side of the threshold we observe $Y(1)$.

2. **Prognosis-weighted difference of intercepts.** Now, we fit regressions to estimate the difference of intercepts of the potential outcomes regression function relating $Y(0)$ to the forcing variable R_i. The regressions (mirroring e.g. De la Cuesta and Imai 2016) are as follows. First,

$$(\widehat{\alpha_0},\, \widehat{\beta_0}) = \underset{\alpha_0,\beta_0}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c_0 \leq R_i \leq c\}\{\widehat{Y_i(0)} - \alpha_0 - \beta_0(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{41}$$

   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). Here, R_i is the forcing variable, c is its value at the assignment threshold, and c_0 is the value that defines the edge of the control-group bandwidth. Similarly,

$$(\widehat{\alpha_1},\, \widehat{\beta_1}) = \underset{\alpha_1,\beta_1}{\arg\min} \sum_{i=1}^{n} \mathbb{I}\{c < R_i \leq c_1\}\{\widehat{Y_i(0)} - \alpha_1 - \beta_1(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{42}$$

   is the intercept and slope from the regression on the treatment-group side, including units up to c_1. As recommended by Calonico et al. (2014) and Cattaneo et al. (2020), equations (35) and (36) 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`.

   For clarity, in (35) and (36), 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 centered value of the forcing variable, $R_i - c$. Note, however, that the latter are distinct from the fitted coefficients of the regression of $Y(0)$ on *covariates* $\boldsymbol{X}$. These, which we label $\widehat{\beta^C}$ as before, are fit in the prognosis regression in the previous step.

Our key test statistic is the difference of prognosis-weighted intercepts of regressions above and below the assignment threshold, i.e., as in text,

$$\delta_{PWLR}^{RD} = \widehat{\alpha_1} - \widehat{\alpha_0}. \tag{43}$$

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 value at which treatment assignment flips from "off" to "on," conditional on the covariates to the right and to the left of the threshold, respectively. Since we do not observe $Y(0)$ on one side of the threshold determining treatment assignment, we take $\widehat{\beta^C}$ from the prognosis regression in the first step (using observations from the control-group side of the threshold). However, the regression on the treatment group side of the threshold uses the treatment group values of $\boldsymbol{X}$ to form the corresponding $\widehat{Y_i(0)} = \boldsymbol{X}\widehat{\beta^C}$ as in equation (13) in the text.

In sum, we regress the fitted $\widehat{Y(0)}$ on the running covariate on each side of the threshold, using a kernel-weighted approach to prioritize units closest to the threshold to fit the intercepts. 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.

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).

Our `R` package inherits the default options for this kernel-weighted local-linear regression from the `rdrobust` package of Calonico et al. (2015). There are three main elements: specification of the bandwidth for the local-linear regression; the kernel function; and the polynomial order of the regression. For the bandwidth, our default uses the MSE-optimal bandwidth of Imbens and Kalyanaraman (2012), symmetric on both sides of the threshold, as in `rdrobust`; here, the optimal bandwidth is selected for the regression of $\widehat{Y(0)}$ on R_i. (Bandwidth selection occurs within the `rdrobust` function, also in the `rdrobust` package, which is a dependency in our package. However, the bandwidth selection function `rdbwselect` of Calonico et al. (2015) can also be called directly). The bandwidth can also be specified manually or chosen using other procedures such as cross-validation available in `rdrobust`. Next, for the kernel, our default uses the triangular kernel noted in the text, but users can specify e.g. a uniform kernel. Finally, the default option uses a polynomial of order 1, i.e., a local-linear regression. All of these default choices can be altered using options inherited from `rdrobust`. For discussion of some of the theory of the kernel-weighted local-linear regressions, see Calonico et al. (2014) and also Cattaneo et al. 2020, chapter 4.

For comparison in our simulations and also to allow researchers the flexibility to consider an unweighted version of the test statistic in RD designs, our `pwtest` also reports a statistic similar to δ_{UW} in tests of as-if random. The test also returns the unweighted sum of the difference of intercepts, estimated separately for each covariate with kernel-weighted local-linear regressions using MSE-optimal bandwidths.

5.1.4 Statistical inference and hypothesis testing

Statistical inference and hypothesis testing is tricky in RD designs—perhaps most fundamentally because it is not always clear what chance process in fact accounts for random variation in the estimators (Bueno

et al. 2014, Cattaneo et al. 2015). In the case of our prognosis-weighted test for continuity of potential outcomes, an additional difficulty is that statistical inference and hypothesis tests must account for random variation not just in estimation of the intercepts of regression functions but also in the prognosis weights. For standard errors, researchers often use model-based solutions or large-sample approximations. Substantial theory on the estimation of standard errors in RD designs has been developed recently; see e.g. Calonico et al. (2014), Calonico et al. (2015), Cattaneo et al. (2015), Cattaneo et al. (2020).

In light of this, we take two complementary approaches to testing, both available in our `R` package `pwtest`.

1. First, we export conventional and bias-corrected standard errors from the `rdrobust` package of Calonico et al. (2015), which users can alternately use for hypothesis testing using normality approximations from large-sample results. This is our default approach. We use it for the p-values reported in Table 1 of the paper (for tests of continuity in RD studies).

2. Second, we also adapt our bootstrap (resampling-based) test in a way that accounts for different sources of chance variation. The test mimics the random variation implied by estimation of $\widehat{\beta^C}$ in the prognosis regression and generates bootstrap bandwidths, in which we estimate intercepts for kernel-weighted local linear regressions. However—precisely because the source of chance variation in RD designs can vary depending on the application, and because of other features of the bootstrap we describe below—we encourage users to study the bootstrap carefully and ensure it matches their application before modifying the default option.

We now discuss these two approaches in more detail.

Approach #1: standard errors and p-values from rddrobust

Our default approach uses estimated standard errors from the `rddrobust` function of Calonico et al. (2015) to form a z-ratio and conduct a test of continuity of average potential outcomes. There are two elements to the z-ratio. In the numerator is the observed difference of intercepts $\delta_{PW}^{RD,obs}$. In the denominator is an estimated standard error for this observed difference of intercepts exported from `rddrobust`. In the function's default, we use the conventional standard errors from `rddrobust` but users can instead use the bias-corrected version (Cattaneo et al. 2020, Calonico et al. 2015). To invoke the default in our package `pwtest`, users specify a test of continuity of potential outcomes in an RD design by setting `rdd = TRUE`. They then set `se_type` to "analytic" (or choose not to specify this option, since it is the default when `rdd = TRUE`).

With a large-sample normality approximation for the difference of intercepts, the ratio of the estimate (difference of intercepts) to the exported standard error for the difference can be referred to a z distribution. Hence, our function estimates the p-value of δ_{PW}^{RS} from a two-tailed test, as follows:

$$p = 2 * P(Z \geq |z|) \tag{44}$$

where $P(\dot{})$ is a probability density function of a normally distributed statistic, z defined as $z = \frac{\delta_{PW}^{RD}}{\widehat{\sigma}_{PW}^{RD}}$ and $\widehat{\sigma}_{PW}^{RD}$ is the conventional (or bias-corrected, as specified by the user in `pwtest`) standard error from a local polynomial RD estimates exported from `rdrobust` (Calonico et al. 2015).

For further theory justifying this approach, see e.g. Calonico et al. (2014) and Cattaneo et al. (2020).

Approach #2: Bootstrap hypothesis test

We also derive a resampling-based (bootstrap) test for δ^{RD}_{PWLR}.

It works as follows. Consider as a test statistic a particular $\delta^{RD,obs}_{PWLR}$, i.e., the observed δ^{RD}_{PWLR} calculated using equations (35)-(36) in section 4.2.1 above with a particular data set.

Then the hypothesis test assesses the statistical significance of $\delta^{RD,obs}_{PWLR}$ in the following steps:

1. **Bootstrap samples.** Suppose there are N_c units on the control group side of the threshold over the support of the forcing variable R_i. For example, if treatment is operative at and above the RD threshold $R_i = c$, N_c is the size of the set of units with $R_i < c$. Draw a sample of size N_c at random with replacement from this population of units $i : R_i < c$ (or $i : R_i > c$, if treatment is operative at and below the threshold). This is the *bootstrap control group*.

   Now, draw another independent sample from the control group units, here however of size N_t (the number of units on the treatment group side of the threshold). This is the *bootstrap treatment group*. The reason for sampling again from the control group is that we want to construct a null distribution of $Y(0)|R_i$ in both the treatment and control groups, and we observe potential outcomes under control only in the control group.

2. **Prognosis regression**. Now, regress $Y(0)$ on covariates in the sample of size $N_c + N_t$. Denote the resulting bootstrap regression coefficient vector $\widehat{\beta^{C*}}$. In the bootstrap treatment group, replace the value $R_i - c$ for each sampled i with $-(R_i - c)$, so that units sampled from below the threshold (when the control group is below the threshold) become units above the threshold, and vice versa. The idea is to "mirror" the absolute value of distance $R_i - c$ of sampled units on each side the threshold, so that those closer to the threshold remain closer to the threshold; in the triangular-kernel weighted local-linear regression we use to fit the bootstrap intercepts in step 4, such units will have a stronger influence on the estimates. Form $\widehat{Y_i(0)}^* = X'\widehat{\beta^{C*}}$ for all sampled i (i.e. pooling the bootstrap treatment and control groups). This is the *bootstrap regression function for potential outcomes under control*.

3. **Bootstrap bandwidths**. Use a bandwidth selector to define the window for the local-linear regressions. Our default uses the MSE-optimal bandwidth of Imbens and Kalyanaraman (2012), symmetric on both sides of the threshold, just as in our calculation of the observed prognostic-weighted difference of intercepts (see step 2 of section 4.2.1 above). Note that here, the bootstrap MSE-optimal bandwidth is selected for the regression of $\widehat{Y(0)}^*$ on R_i, i.e., the regression of the bootstrap predicted value on the forcing variable. (Users can alter this bandwidth manually by passing an optional argument `h` to our `pwtest` function call, which passes it on to Calonico et al. (2015)'s `rdbwselect` function).

   This produces a bootstrap bandwidth $[c^*_0, c^*_1]$, with $c^*_0 < c < c^*_1$. Note that the bandwidth sizes are realized values of random variables, as they are functions of $\widehat{Y(0)}^*$ in the previous step. Let n^*_0 be the size of the set of sampled units $i : R_i \in [c^*_0, c)$ and correspondingly n^*_1 is the size of the set of units $i : R_i \in [c, c^*_1]$; if treatment is operative at and above the threshold, these are control and treatment-group units, respectively. (Conversely, if treatment is operative at and below the threshold, switch the labels n^*_1 and n^*_0). The set of units within the bootstrap bandwidth $[c^*_0, c^*_1]$ thus has size $n^* = n^*_0 + n^*_1$.

4. **Prognosis-weighted difference of intercepts.** Now, conduct bootstrap regressions that mimic equations (12) and (13) in the text. For instance, if treatment is operative below the threshold, we have

$$(\widehat{\alpha_0}^*, \widehat{\beta_0}^*) = \underset{\alpha_0,\beta_0}{\arg\min} \sum_{i=1}^{n^*} \mathbb{I}\{c_0^* \leq R_i \leq c\}\{\mathbf{X_i}'\widehat{\beta^{\mathbf{C}*}} - \alpha_0 - \beta_0(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{45}$$

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

$$(\widehat{\alpha_1}^*, \widehat{\beta_1}^*) = \underset{\alpha_1,\beta_1}{\arg\min} \sum_{i=1}^{n^*} \mathbb{I}\{c < R_i \leq c_1^*\}\{\mathbf{X_i}'\widehat{\beta^{\mathbf{C}*}} - \alpha_1 - \beta_1(R_i - c)\}^2 K\left(\frac{R_i - c}{h}\right) \tag{46}$$

is the intercept and slope from a regression of $\widehat{Y(0)^*}$ on the running covariate to the right of the threshold (centered at the threshold). This results in the bootstrap intercepts $\widehat{\alpha_0}^*$ and $\widehat{\alpha_1}^*$. For these bootstrap regressions, use whatever options are specified for the placebo test regression using the observed data—e.g., triangular kernel, polynomial of order 1, as in our default.

Putting together the two regressions allows to form a bootstrap placebo treatment effect estimator $\delta_{PWLR}^{RD*} = \widehat{\alpha_1}^* - \widehat{\alpha_0}^*$

5. Repeat steps (1)-(2) B times (we set $B = 500$ in our default).

6. We then calculate a two-sided randomization-based p-value as

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(|\delta_{PW,b}^{RD*}| \geq |\delta_{PWLR}^{RD,obs}|), \tag{47}$$

where $\mathbb{1}$ is an indicator function that takes on the value of 1 if its argument is true and 0 otherwise.

Here $\delta_{PWLR}^{RD,obs}$ is the observed prognosis-weighted difference of intercepts and $\delta_{PW,b}^{RD*}$ is the bth bootstrap prognosis-weighted difference of intercepts. Reject the null if, say, $p^* < 0.05$.

Thus, in this approach, we compare the absolute value of the observed value of the prognosis-weighted difference of intercepts to its randomization distribution and reject the null hypothesis if such an observed value would arise in fewer than 5% of randomizations under the null.

To use the bootstrap (resampling-based) hypothesis test for the continuity of average potential outcomes, users first specify an RD design by setting `rdd = TRUE`) in `pwtest`). They then set `se_type` to "bootstrap."

The bootstrap is useful in some contexts but we opt to use the `rdrobust` options as the default for the following reasons. First, note that one feature of the bootstrap prognosis regression in step 2 is that it imposes a particular functional form on the regression functions on each side of the threshold, because the treatment group values of the centered running variable "mirror" those on the control group side. That, since we form $\widehat{Y_i(0)}^* = X'\widehat{\beta}^{C*}$ for all sampled i and a sampled unit used for the treatment ground regression has only its $R_i - c$ value altered (by changing its sign), the value of $\widehat{Y_i(0)}^*$ given X and $\widehat{\beta}^C$ for a given i is the same whether in treatment and control. If the control-group regression is upward-sloping

and the control group is to the left of the threshold, this will produce an "inverted V" shape to the fitted potential outcomes regression function. Clearly, this introduces an assumption on the null distribution of the function that is stronger than just continuity at the threshold—one reason we do not make the bootstrap our default approach for testing the continuity of potential outcomes in RD designs.

However, we also note that this may be a feature as much as bug. We care mainly under this approach that the function is continuous at the threshold; use of the kernel-weighted local-linear regression will give maximal weights to units closest to the threshold, so the shape of the regression functions far from the threshold may be less relevant. Moreover, this choice (or a similar one) appears unavoidable, since we need to simulate a null distribution that gives continuity at the threshold and thus do not want to sample X values from the observed treatment group (since in the data set at hand, continuity may not hold). We thus here opt for this approach in the bootstrap. Nonetheless, researchers should bear in mind possible implications of these features of the bootstrap in the context of their applications. Future research may consider additional modifications to the options available currently in the package.

5.2 Testing as-if random in RD designs

Notwithstanding the primary role of continuity, in some regression-discontinuity (RD) designs, it may be appropriate and desirable to test an assumption of as-if random assignment in a small neighborhood of the threshold determining treatment assignment. This may especially be so when the regression functions relating potential outcomes to the running covariate are flat (formally, the first derivative of the functions are near zero; see e.g. Dunning 2012 Chapters 3 and 5).[25] The shape of the potential outcome regression functions can be assessed by, for instance, fitting flexible regressions on each side of the threshold determining treatment assignment. This effectively treats the RD design as if it might be a locally randomized experiment (Lee 2008, Lee and Lemieux 2010; also Sekhon and Titiunik 2017).

The techniques we develop in sub-sections 4.1 and 4.2 of the paper readily apply to a test of as-if random in an RD design. Within a given window, the relevant null hypothesis is H_0 in (4) in the text (or equivalence testing can be used). The control group consists of units whose score (running covariate) is above or below the relevant threshold, depending on where treatment is operative, within the window.

The researcher then fits the standardized regression of outcomes on covariates in the control group to estimate the vector β; and she calculates standardized differences of treatment and control group means to form δ_{PWLR} in equation (7), i.e., the weighted sum of the differences of means. For hypothesis testing, our resampling-based hypothesis test readily applies: when the RD design creates a locally randomized experiment, treatment and control groups are exchangeable. Our bootstrap simulates the null distribution of our test statistic to which we may compare the observed δ_{PWLR} to find p-values.

The major challenge is selection of the window or bandwidth within which as-if random is plausible and should be tested. For example, Cattaneo et al. (2015) develop hypothesis tests for the RD design based on randomization inference, relying on an assumption of "as-if" random assignment within a narrow window around the key threshold. For placebo and balance tests, the major difficulty is that there is some apparent circularity: the bandwidth should be chosen so that as-if random holds, but the balance tests are supposed to tell you if as-if random holds. Thus, it is not necessarily obvious how to choose the bandwidth for purposes of testing.

[25] As-if random also motivates exact statistical tests based on randomization inference, including those proposed by Cattaneo et al. (2015); these do not rely on large-sample approximations and thus may be especially helpful in small studies or when data are sparse near the RD threshold.

At least two approaches are plausible. One is the sequential testing procedure described by Cattaneo et al. (2015), in which a researcher iteratively shrinks the window until she is unable to reject the null of as-if random. Cattaneo et al. (2015) describe a randomization inference approach to testing. As they note (see also Imai et al. 2008), this procedure is vulnerable to Type II error (failure to reject the null when it is false) due to a small number of units with the window at which testing stops. Equivalence testing could therefore be one alternative, using the same type of sequential test though potentially *increasing* the window until the null of difference can be rejected. Note that whereas the sequential procedure described by Cattaneo et al. (2015) does not describe how to adjudicate across tests with different pre-treatment covariates (which may generate different maximum window sizes for different covariates), our omnibus test using δ_{PWLR} will result in a single window, because it is based on a single regression.

A second, related possibility is to present results of tests of as-if random (i.e. tests of H_0 in equation 2 in the text) for a large variety of bandwidths, using a graphical procedure like that described by Bueno and Tuñón (2015). This approach can also be viewed as sequential, in the sense that one can identify visually the largest region in which tests do not reject as-if random. However, it allows a stronger visual sense of how test results hold across varied potential bandwidths. Again, the procedure could be subject to the "balance test fallacy" (Imai et al. 2008), in that tests with smaller study groups are less well-powered to reject the null and the study group size decreases as the bandwidth shrinks. Equivalence testing can address this concern, but the need to specify an equivalence range remains a drawback of the approach (Hartman and Hidalgo 2018). Instead, one could inspect the confidence intervals for different bandwidth sizes and, for the smallest bandwidth in which as-if random is not rejected, consider the range of imbalances contained within the interval, using either a traditional or equivalence testing approach.

The advantage of our test, with either approach to defining the bandwidth, is that as elsewhere, assessment will be based on the most prognostic covariates and thus will allow a readier test of as-if random, i.e., the independence of treatment assignment and potential outcomes.

5.2.1 The relationship between running variables and outcomes in sampled RD studies

Figure A1 shows the relationship between the outcome and running variables around the cutoff point for all studies in our sample that employ a regression discontinuity design. The plots are generated using the `rdplot` function in the `rdrobust` R package. We use the function's default binning method, which mimics variance evenly spaced using spacings estimators (see Calonico et al. 2014). The discontinuity sample for each study is defined according to the bandwidths described in the previous section.

We specify the linear fit in order to examine whether the relationship between the outcome variable and running variables is flat, which may suggest an argument for using the as-if random test in these settings. Our prognosis-weighted test focuses on potential outcomes under control, so the control group side of the threshold is most relevant for comparing the results of our tests to the shape of the potential outcomes function. Note that the prognosis-weighted tests will be most specific and powerful when the prognosis R^2 is high. (If it is not, the test may not well approximate the potential outcomes under control depicted graphically on the control group side of the threshold in the plots).

Notice that per Table 1 in the paper, we fail to reject as-if random with a prognosis-weighted test in 7 studies, while we reject it in 5 studies. The studies in which as-if random is not rejected are: (1) Hall (2015); (2) Kim (2019); (3) Novaes (2018); (4) Hidalgo and Nichter (2016); (5) Fournaies and Hall (2014); (6) Boas and Hidalgo (2011); (6) Eggers et al. (2015).

Conversely, the studies in which as-if random is rejected in the prognosis-weighted test are: (1)

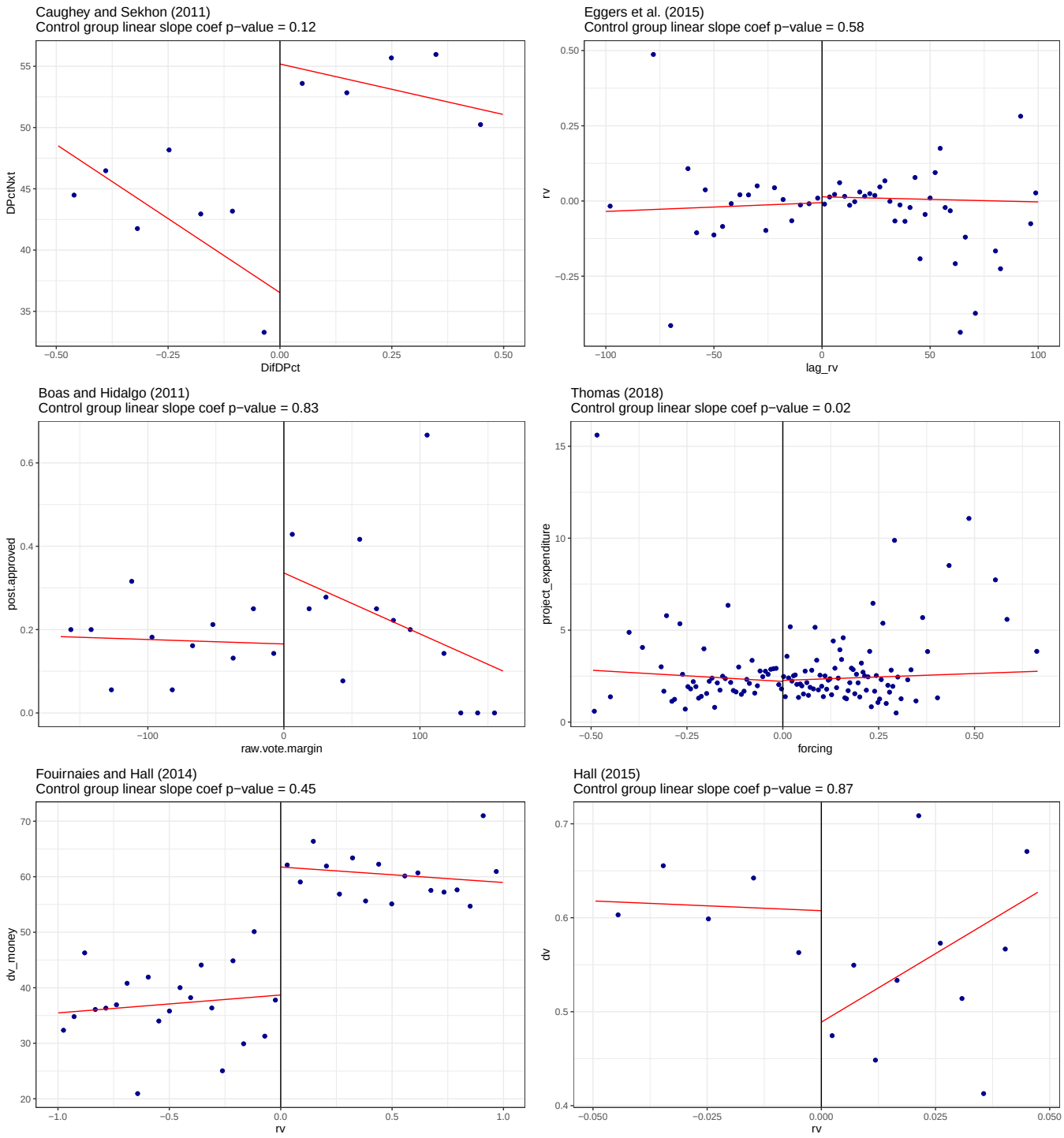
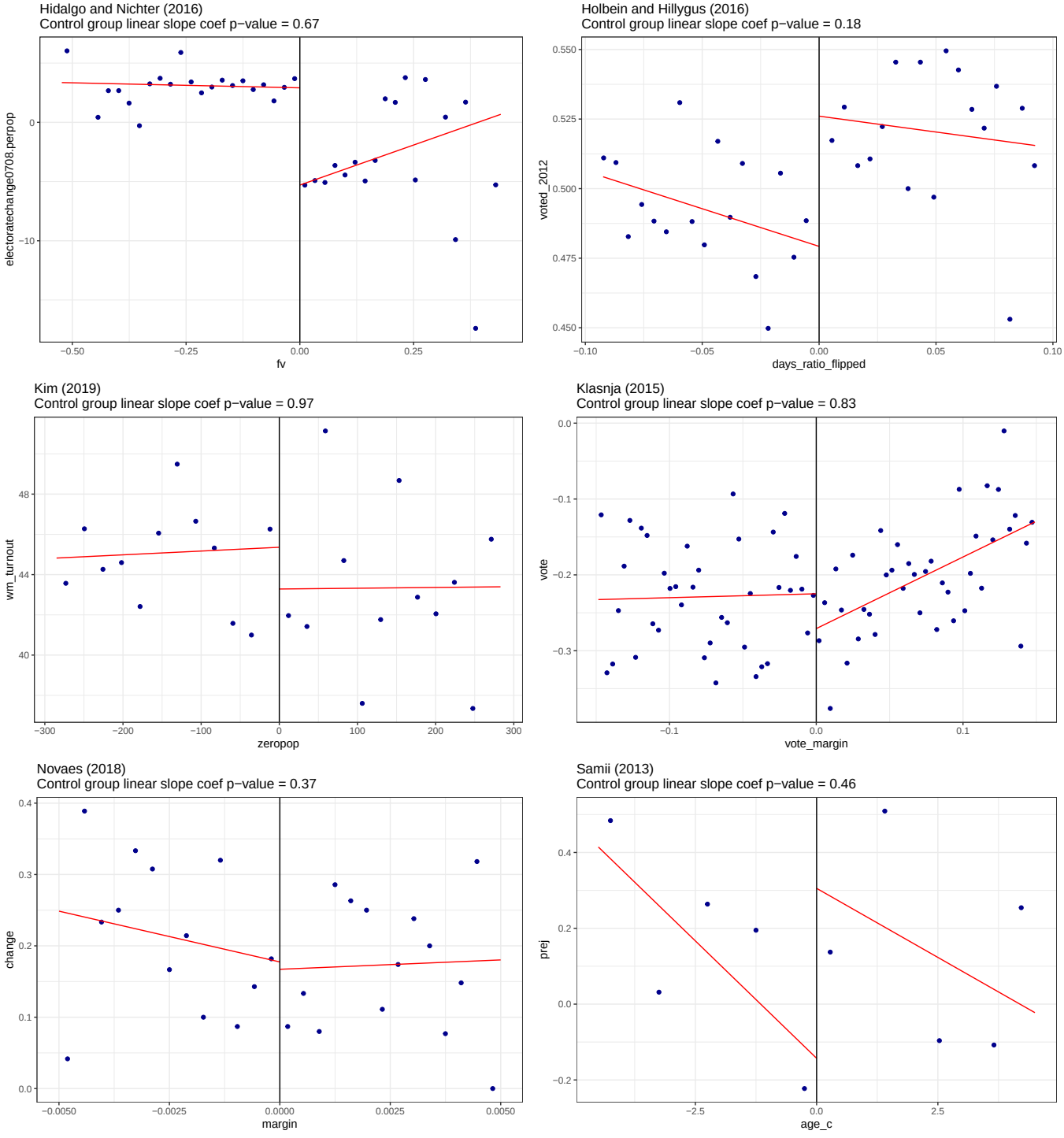


Figure A1: Outcome over binned running variable values within the RD bandwidth defined for each study in the previous section. Red line shows the linear regression line on either side of the threshold.


Hidalgo and Nichter (2016)
Control group linear slope coef p−value = 0.67
electoratechange0708.perpop
fv
Holbein and Hillygus (2016)
Control group linear slope coef p−value = 0.18
voted_2012
days_ratio_flipped
Kim (2019)
Control group linear slope coef p−value = 0.97
wm_turnout
zeropop
Klasnja (2015)
Control group linear slope coef p−value = 0.83
vote
vote_margin
Novaes (2018)
Control group linear slope coef p−value = 0.37
change
margin
Samii (2013)
Control group linear slope coef p−value = 0.46
prej
age_c

(Fig A1 cont.)

Caughey and Sekhon (2011); (2) Samii (2013); (3) Thomas (2018); (4) Kasnja (2015); (5) Holbein and Hillygus (2016).

It is suggestive that visually, those RD studies where the regression function for potential outcomes under control appears flattest are those in which we fail to reject as-if random: Eggers et al. (2015), Boas and Hidalgo (2011), Founaies and Hall (2014), Hall (2015), Hidalgo and Nichter (2016), Kim (2019), and Novaes (2018). An exception is Thomas (2018), where the slope appears flat but the prognosis-weighted test rejects as-if random in Table 1. Conversely, the slopes are steeper in those studies prognosis-weighted tests reject as-if random: see especially Caughey and Sekhon (2011), Samii (2013), and Holbein and Hillygus (2016).

While not the main focus of our paper, future work should explore this empirical variation across different types of RD studies in the relationship between the running covariate and potential outcome regression functions.

6 Prognosis-weighted equivalence tests

Prognosis weighting can also be adapted to take advantage of equivalence tests (Hartman and Hidalgo 2018). Equivalence tests seek to address the "balance test fallacy" (Imai et al. 2008, Section 7), in particular, the problem that failing to reject the null of as-if random is not the same as accepting it. With traditional tests, researchers may fail to reject simply because a study is small and underpowered.

The test works by switching the null and alternative hypotheses, so that under the null, the expected means in the treatment and control group differ, while under the alternative they are approximately equal. Equivalence tests are less likely to reject the null of difference as study size shrinks (Hartman and Hidalgo 2018, Figure SI-2), so acceptance (rejection of the absence) of as-if random is less likely to be an artifact of low power.

Prognosis-weighted equivalence tests can provide an additional protection against the balance test fallacy. In Online Appendix Section 5, we adapt the bootstrap procedure in subsection 4.4 for equivalence testing. Here, the most informative covariates must be sufficiently balanced to reject the null hypothesis of difference. Thus, as long as covariates are sufficiently jointly informative, prognosis weighting ensures that we will not "accept" as-if random unless covariates related to potential outcomes are sufficiently balanced.

It is important to note, however, that an equivalence test based on covariates with weak joint prognosis is subject to similar limitations as traditional tests. Thus, we may reject the absence of as-if random based on the balance of the most prognostic individual covariates, among the set at our disposal. Yet, if measured covariates are not as a whole prognostic, there could readily be lurking prognostic variables that are unobserved and imbalanced. Were we successfully to measure these prognostic covariates, we might instead reject (fail to reject the absence of) as-if random.[26]

The way around this difficulty—as with traditional testing—is to ensure that we have measured covariates that are adequately jointly prognostic. The best advice may be thus to develop high-powered tests—either traditional or equivalence-based—by leveraging jointly prognostic covariates and then prioritizing balance of the most informative individual covariates, as in our prognosis-weighted test.

[26] A further drawback is that researchers may find evidence for or against as-if random by varying the equivalence 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.

6.1 A bootstrapped equivalence test p-value

Here, we adapt the bootstrap procedure in subsection 4.1.1. in the paper for equivalence testing.

The key idea in the equivalence test is that under the null hypothesis, the treatment and control groups are drawn from different distributions—and thus, e.g., as-if random fails—whereas under the alternative, they are drawn from approximately equivalent distributions. For as-if random, the null hypothesis is that the assumption does *not* hold, while the alternative is that expected values of average potential outcomes in the two groups are approximately equal:

$$\begin{aligned} H_{0_{\text{equiv}}} &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] \neq 0 \\ H_{A_{\text{equiv}}} &: \mathbb{E}[\overline{Y(0)^T} - \overline{Y(0)^C}] \approx 0. \end{aligned} \tag{48}$$

What constitutes approximate equality under the alternative is captured, in a "two one-sided test" (TOST), by the "equivalence range," i.e., the requirement that $E[\overline{Y(0)^T} - \overline{Y(0)^C}]$ be contained in some interval $[\epsilon_L, \epsilon_U]$.

Adapting our notation in section 4.1.2 of the paper, we can state

$$H_{0_{\text{equiv}}} \quad : \quad \mathbb{E}[\frac{\overline{Y(0)^T} - \overline{Y(0)^C}}{\sigma}] \geq \epsilon_U \quad \text{OR} \quad \mathbb{E}[\frac{\overline{Y(0)^T} - \overline{Y(0)^C}}{\sigma}] \leq \epsilon_L \tag{49}$$

versus

$$H_{A_{\text{equiv}}} \quad : \quad \epsilon_L < \mathbb{E}[\frac{\overline{Y(0)^T} - \overline{Y(0)^C}}{\sigma}] < \epsilon_U.$$

where $\epsilon_L < 0 < \epsilon_H$ and σ is the standard deviation of the potential outcomes under control in the finite population. Thus, under the null, the standardized difference between the mean of the treatment and control group distributions is greater than the positive upper bound ϵ_U or less than the negative lower bound ϵ_L. Under the alternative, it lies within this range (Hartman and Hidalgo 2018: 1003).

The equivalence range may be chosen using substantive knowledge of what would constitute trivial differences under the alternative (perhaps relative to a treatment effect size estimated in previous studies). As a default, Hartman and Hidalgo (2018, 1006) recommend $[\epsilon_L, \epsilon_H] = \pm 0.36\sigma$.

We can devise a randomization inference test of the null hypothesis in (49) using the union of two one-sided exact tests. Appealing to the intersection-union principle and following Hartman and Hidalgo (2018, 1009), we "conduct one-sided tests of the strict null hypothesis equal to the bounds of the equivalence range, and the overall null hypothesis of nonequivalence can be rejected if both corresponding permutation p-values are less than the level of the test." Thus, we use the intersection of the p-values for the two one-sided tests to define the rejection rule.

The key difference in the bootstrap routine—relative to the procedure given in subsection 4.1.1 of our paper for a traditional test—is that we must draw the treatment group from a different distribution from the control group. Thus, in one one-sided test, we draw the treatment group from a distribution centered on the upper bound of the equivalence range. In the other one-side test, the distribution is centered on the lower bound of the range. To accomplish this, we add or subtract (depending on whether we are doing one-sided tests for ϵ_H or ϵ_L) a fixed value for each control group observation, before sampling the treatment group. See Arboretti et al. (2018, 9-10) for a related algorithm that, however, is based on a permutation of

treatment assignment. Permutation is not viable in our case, as explained in the text, since we do not wish to mix values of $Y(0)$ and $Y(1)$, so we use a resampling-based randomization inference approach instead.

We can use the default fixed value of 0.36σ recommended by Hartman and Hidalgo (2018) to define the bounds of the equivalence range. The resampling-based equivalence test works as follows.

One-sided test for ϵ_H:

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. Now, **add 0.36σ to the values for each control group observation**. Then, sample treatment group values independently from this modified bootstrap population to calculate $\overline{X^{T*}}$ and the simulated $\delta_{PW}^{*b,\epsilon_U} = (\overline{X^{T*}} - \overline{X^{C*}})'\widehat{\beta}^{C*}$. (We use $\delta_{PW}^{*b,\epsilon_H}$ to denote that this is the bootstrap $\delta_P W$ when drawing from a distribution centered on the upper boundary of the equivalence range, ϵ_H).

3. Repeat steps (1)-(2) B times ($B = 500$ in our default).

4. Calculate a one-sided randomization-based p-value as

$$p_{\epsilon_U}^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\delta_{PWLR}^{*b,\epsilon_U} \geq \delta_{PWLR}^{\text{obs}}), \tag{50}$$

   where $\mathbb{I}$ is an indicator function that takes on the value of 1 if its argument is true and 0 otherwise.

   Note that unlike in the two-sided traditional test given in section 4.1.1, here there are no absolute value symbols around δ_{PWLR}^{*b} and $\delta_{PWLR}^{\text{obs}}$: we are conducting a one-sided test of the null that the difference of expectations in (49) is greater than or equal to ϵ_U.

One-sided test for ϵ_L:

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. Now, **subtract 0.36σ from the values for each control group observation**. Then, sample treatment group values independently from this modified bootstrap population to calculate $\overline{X^{T*}}$ and the simulated $\delta_{PW}^{*b,\epsilon_L} = (\overline{X^{T*}} - \overline{X^{C*}})'\widehat{\beta}^{C*}$. (We use $\delta_{PW}^{*b,\epsilon_L}$ to denote that this is the bootstrap $\delta_P W$ when drawing from a distribution centered on the lower boundary of the equivalence range, ϵ_L).

3. Repeat steps (1)-(2) B times ($B = 500$ in our default).

4. Calculate a one-sided randomization-based p-value as

$$p_{\epsilon_L}^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\delta_{PWLR}^{*b,\epsilon_L} \leq \delta_{PWLR}^{\text{obs}}). \tag{51}$$

   Rejection rule: For a 0.05-level test, reject the null of difference if both $p_{\epsilon_U}^*$ AND $p_{\epsilon_L}^*$ are less than 0.05.

In words, we reject the absence of as-if random if the observed prognosis-weighted test statistic is statistically less than the upper bound of the equivalence range and statistically greater than the lower bound of the equivalence range.

7 Performance of prognosis-weighted tests: Simulations

In this section, we turn to simulations to assess the performance of our informativeness-weighted tests. The structure of the simulations allows us to assess the power and specificity of the tests (i) when as-if random is true, that is, prognostic covariates are balanced in expectation; and (ii) when it is false, that is, treatment assignment depends on potential outcomes.

In a first set of simulations (subsection 7.2), measured covariates are jointly fully informative (i.e., sufficient) but vary in their individual prognosis. Thus, we assess how control over Type I and Type II error (false positives and false negatives) of different types of tests responds to changes in the prognosis of covariates that are either balanced or imbalanced in expectation.

In a second set (subsection 7.3), covariates are fully or partially uninformative, and we assess the performance of tests as covariates become more prognostic. This structure also permits analysis of threshold levels of covariate prognosis at which, in the simulations, performance becomes adequate, for instance, at which the tests achieve specified statistical power (subsection 7.4).

In these first two sets of simulations, we compare the performance of the baseline prognosis-weighted test based on linear regression to that of two unweighted multivariate test statistics: the unweighted sum of standardized covariate differences of means (δ_{UW} in subsection 4.3.1) and Hotelling's T^2. For simplicity, here potential outcomes are also a linear function of covariates in the data-generating process.

In a final set of simulations (subsection 7.5), we then relax both the linearity of the data-generating processes and of the prognosis-weighted tests. Thus, in subsection 7.5.1, potential outcomes are a polynomial function of covariates, and we compare unweighted and baseline prognosis-weighted tests to a regression-based test that allows polynomial functions of the covariates. In subsection 7.5.2, we further allow covariate interactions both in the data-generating process and in the calculation of the test statistics.

Last, in subsection 7.5.3, we evaluate a range of further non-linear tests, including those based on machine learning, in the presence of two 'difficult', highly non-linear data-generating processes. Here, we also assess our tool for automated selection of a test procedure based on the best-fitting (regression or machine-learning) technique for predicting $Y(0)|X$ in the control group. Thus, we compare the performance of the "winner" (using the `contest` and `pick_winner` functions in our `pwtest` package) to that of specified tests, including those based on linear and saturated linear (i.e., fully polynomial and interactive) regressions.

In all the simulations, we compare the performance of tests given different expected patterns of imbalance on prognostic and non-prognostic covariates. We measure performance as the proportion of rejections of as-if random across realizations of a particular d.g.p and treatment assignment vector. When at least one of the prognostic variables is imbalanced in expectation, treatment assignment depends on potential outcomes so as-if random fails.[27] In this case, the rejection rate measures statistical power. In contrast, when as-if random holds, the proportion of rejections measures the false positive rate (Type I error).

Overall, our results illustrate how prognosis weighting can reduce both false negatives and false positives. The performance of the tests depends on the overall prognosis of measured covariates. In contrast,

[27] We implement a minor technical correction to ensure this holds; see our replication code.

unweighted tests that do not use information on covariate prognosis sacrifice power and/or specificity. Further gains in power can sometimes be achieved through the use of non-linear tests, but this depends not only on the relationship between covariates and potential outcomes but also on patterns of imbalances, for instances, whether main (linear) or secondary (non-linear) terms are relatively imbalanced. Overall, the linear test performs quite well, especially when it is is expanded in saturated form to include polynomials and covariate interactions. We return to the main takeaways from the simulations in subsection 7.6.

7.1 Steps in the simulations

Our simulations proceed in the following steps:

Step 1: (Data-generating process). We generate a dataset of $N = 500$ observations. The dataset has a treatment assignment vector Z (with half the units assigned at random to treatment and half to control); potential outcomes $Y(1)$ and $Y(0)$; and covariates X_p, with $p = 1, 2$ or sometimes $p = 1, 2, 3$. Covariates are drawn from a multivariate normal distribution with mean 0 and standard deviation 1, and the elements of the variance-covariance matrix governing the variables are defined in such a way that the expected correlation between covariates and treatment assignment Z is determined by that covariate's imbalance parameter; the expected correlation between covariates is 0. For data-generating processes with interaction terms, the generation of imbalance and independence of covariates is further described in subsection 7.5.2. Potential outcomes under control are formed as a linear or non-linear function of covariates. In the first set of simulations, $Y(0) = \beta_1 X_1 + \beta_2 X_2$, where β_1 and β_2 determine the prognosis of the corresponding covariate. Data-generating processes for potential outcomes in other simulations are described further bleow. The average treatment effect is zero throughout, i.e. $Y_i(0) = Y_1(1)$ for all i, but this plays no role in the simulation. The data-generating processes therefore allow us to set a priori values of covariate imbalance and prognosis, which will be useful in Step 5 below.

Step 2: (Observed test statistics). Using the covariate values in the realized treatment and control groups and the observed $Y(0)$ in the control group, in the dataset generated in Step 1, we calculate test statistics appropriate for the particular technique being evaluated (e.g., δ_{UW}, δ_{PWLR}, or Hotelling's T^2 for linear tests; expanded versions of δ_{PWLR} with polynomial or interactive terms, or statistics based on gradient boosted trees and random forests, for non-linear tests.[28]

Step 3: (Resampling test). We conduct the resampling-based hypothesis tests described in subsection 4.2 in the paper (with $B = 500$). Thus, we calculate p-values appropriate for the particular observed test statistic (e.g. from equation (51) for δ_{PWLR}; we also calculate analogous randomization p-values for δ_{UW}, Hotelling's T^2, and other statistics). Thus, we compare the "observed" test statistics from Step 2 to their randomization distributions when treatment assignment is statistically independent of potential outcomes, as well as covariates. We reject the null hypothesis of as-if random when $p^* < 0.05$.

Step 4: (Rejection rates). We repeat Steps 1-3 1000 times for a given expected correlation structure. That is, on each of the 1000 runs, we produce a dataset of $N = 500$ observations with the given expected covariate imbalance and prognosis. From this, we can calculate the *rejection rate* of each test: the proportion of rejections across the 1000 runs.

[28] We use the `Hotelling` package in R.

Step 5: (Varying prognosis and imbalance). We repeat steps 1-4 with different parameter values determining covariate imbalance and prognosis.

Step 6A: (Minimal sufficiency, sufficiency, and not sufficiency). The first two sets of simulations explore the following situations:

Case 1 (Minimal Sufficiency). In one set of simulations (comprising Steps 1-5), observed covariates are minimally sufficient: in each data set generated in Step 1, $Y(0)$ is a (linear) function of standardized X_1 and X_2, and we use X_1 and X_2 in the observed treatment and control groups to form δ_{PWLR}, δ_{UW}, δ_{PWLR}, and Hotelling's T^2.

Case 2 (Sufficiency). In another set, the observed covariates are sufficient but not minimally so: again, $Y(0)$ is a (linear) function of standardized X_1 and X_2 but we use the observed X_1, X_2, and X_3, where X_3 is a random variable taken from a standard normal distribution. X_3 is unrelated to potential outcomes but may be related to Z, i.e., imbalanced. This case captures the presence of an irrelevant covariate in the test of as-if random.

Case 3 (Not Sufficiency). Finally, we consider a set of simulations in which observed covariates are not sufficient: again, $Y(0)$ is a (linear) function of standardized X_1 and X_2 but we observe only X_1 and X_3, where X_3 is defined the same way as in Case 2. We vary the prognosis and imbalance of X_1 as well as the imbalance of X_3 and fix the prognosis and imbalance of the unobserved covariate X_2. Specifically, we fix X_2 prognosis at 0.25 and set $cor(X_2, Z) = 0.15$.

Step 6B (Prognosis R^2 parameter). In the final set of non-linear simulations, we generate potential outcomes as a non-linear function of covariates but introduce noise such that the overall prognosis of the covariates is governed by a single parameter λ. See section 7.5.

Simulations were run on the High Performance Computing (Savio) server at the University of California, Berkeley. The process outlined in Steps 1-6 runs in parallel on 24 CPU and takes on average 40 hours.

7.2 Informative covariates

We first consider cases in which measured covariates are fully informative, i.e., sufficient. In a first set of simulations, we observe only signal covariates associated with potential outcomes, so the covariates are minimally sufficient. In a second set, we also measure a noise variable X_3 that is imbalanced but is not prognostic; here, covariates are thus sufficient but not minimally so. Thus, we can assess whether a prognosis-weighted test can mimic results from a minimally-sufficient set of prognostic covariates by projecting out irrelevant covariates. The degree of prognosis varies across different simulations.

Figure A2 shows results for one set of simulations. In each plot, the prognosis (true standardized coefficient) of covariate X_1 varies along the vertical axis, while the vertical axis measures the proportion of the 1000 tests in which the null is rejected. The four plots depict results with varied imbalance (measured as the expected correlation ρ between the covariate and treatment assignment) of the potentially prognostic variable X_1 and the noise covariate X_3. A correlation of $\rho = 0.1$ between a covariate and treatment assignment corresponds to an expected standardized difference of means across treatment and control groups of about 0.2. The variable X_2 is set to have a fixed prognosis of 0.25 and is always balanced in expectation.

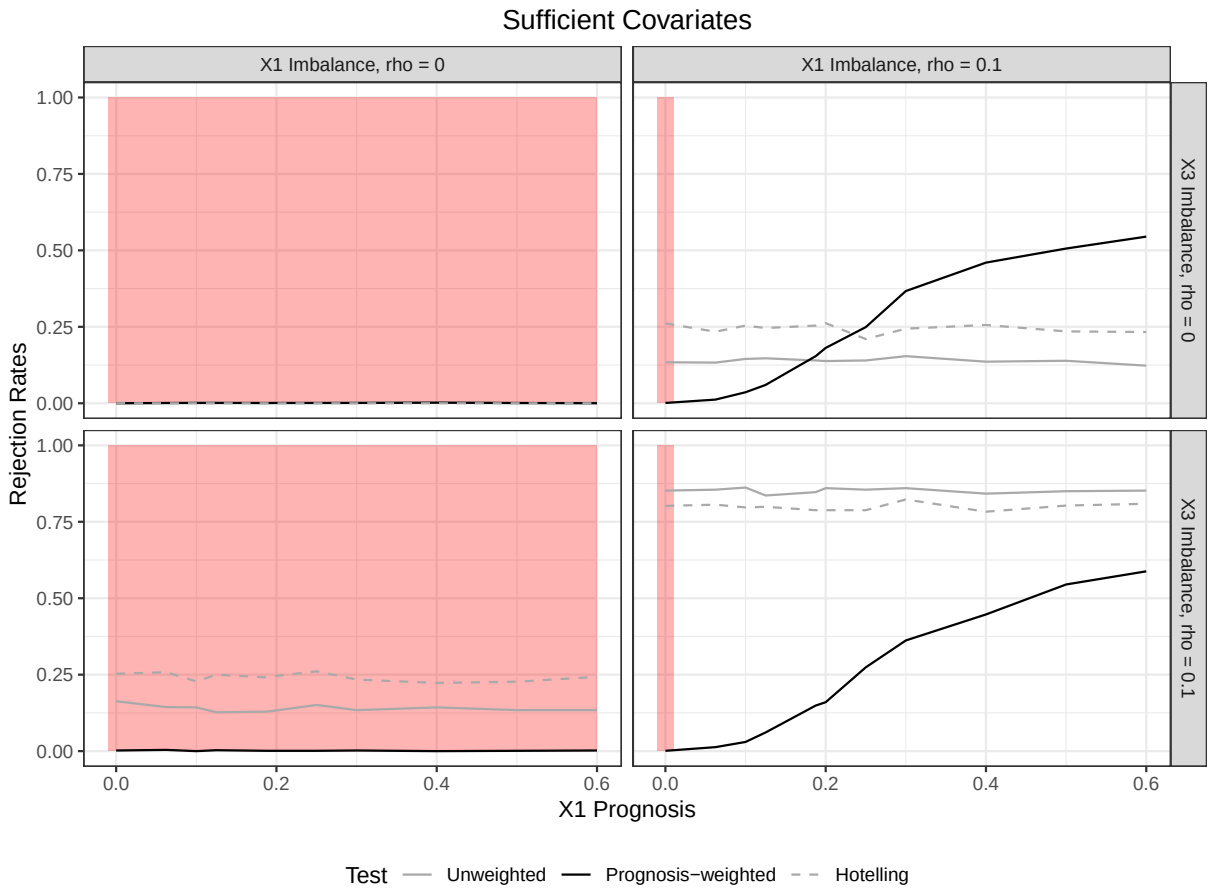


Figure A2: Informative covariates

The figure plots rejection rates as a function of X_1 prognosis and X_1 and X_3 imbalance in a simulation, for prognosis-weighted (dark solid line) as well as unweighted tests. Shaded areas are parts of the parameter space in which as-if random holds. See the text in section 7.2 of this Appendix for further details.

In the top-left panel, all variables are independent of treatment assignment so as-if random everywhere holds, regardless of the prognosis of X_1 (as indicated by the pink shading). Here, the weighted and unweighted tests both perform well, controlling Type I error at similar rates. In the top-right panel, by contrast, as-if random holds only at the origin, when the prognosis of X_1 is zero. Away from the origin, X_1 is both prognostic and imbalanced in expectation, so as-if random fails.

Two aspects of the results in the top-right panel are noteworthy. First, as indicated by the flatness of the dashed and light solid lines, the unweighted tests are completely insensitive to the relative prognosis of covariates. They thus reject at similar rates, whether X_1 is prognostic or not—and thus whether as-if random is true or false. Second and in contrast, the informativeness-weighted test is sensitive to prognosis: as X_1 prognosis increases, the test increasingly prioritizes the imbalance on X_1 and rejects at rising rates. Thus, in the top-right panel we see the prognosis-weighted test achieving both specificity (limiting Type I error when as-if random is true) and power (limiting Type II error when it is false).

We see a similar pattern—insensitivity of unweighted tests to prognosis and increasing power of the

informativeness-weighted test as prognosis increases—when we add imbalance of the noise covariate X_3 (bottom panels). Now, however, the unweighted tests incorrectly reject as-if random at higher rates when it is true (bottom-left panel, and at the origin in the bottom-right panel). This is due to their sensitivity to the imbalance of the irrelevant covariate X_3. The prognosis-weighted test, by contrast, projects out X_3 and bases the test solely on the prognostic covariates, which are balanced in the left panel. Thus, it has a lower false positive rate than the unweighted tests.

As for the lower-right plot, we again see the prognosis-weighted test boosting both specificity and power: when as-if random holds, at the origin with zero X_1 prognosis, it correctly fails to reject the null, whereas rejection rates are increasing in the prognosis of X_1. By contrast, unweighted tests—sensitive as they are to the imbalance of the noise covariate X_3—reject the null at elevated rates whether it is true or false. Although for some parameter values unweighted tests correctly reject as-if random when it is false at higher rates, they also incorrectly reject as-if random when it is true. Because the relative prognosis of covariates plays no role, the unweighted tests do not balance specificity—failing to reject as-if random when it is true—and power—correctly rejecting as-if random when it is false.

Table A3 reports numerical results depicted graphically in Figure A1.

7.2.1 Full set of simulations with informative covariates

While Figure A2 and Table A1 report results from one set of simulations, we reach similar conclusions from a broader set of simulations with greater variation in both imbalance and prognosis parameters.

In Figure A3, we compare the rejection rates of the prognosis-weighted test and Hotelling's T^2 when the null is true and when it is false, across this broader set. Here, we compare rejection rates as a function of the imbalance R^2 (calculated from the regression of treatment assignment on covariates) and the overall prognosis R^2 (calculated from the regression of control outcomes on covariates). When as-if random is true (right plot), the tests offer similar control over Type I error except when prognosis is low, in which case the unweighted tests overreject due to their sensitivity to imbalances on non-prognostic covariates. When the null is false, however, the relative power of the prognosis-weighted test increases with the prognosis R^2. This is particularly important when imbalance is modest but prognostic covariates are imbalanced.

Thus, in Figure A3, for a given set of simulations with particular expected correlations, we plot the balance R^2—that is, the average R^2 from the regression of treatment assignment on all relevant covariates for each case, across all the simulations–against the prognosis R^2, or the average R^2 from the regression of potential outcomes in the control group on all covariates. Thus, we put the simulation results in the same imbalance-prognosis space as in Figure 1 in the paper. We code the simulations captured by each data point according to whether the prognosis-weighted test rejects with greater probability (black points), the unweighted test rejects with greater probability (red points), or the tests reject at the same rate (grey points).

As with Figure A2, in Figure A3 we consider two situations: either as-if random is false (left panel) or as-if random is true (right panel). In the left panel, when the prognosis R^2 is near zero, the unweighted tests often reject at higher rates, sensitive as they are to imbalance in non-prognostic covariates; this occurs especially at very high levels of imbalance. When prognosis is more substantial, the tests reject at equal rates when imbalance is also substantial—often reflecting the patterns in the simulations in Figures 1 and 3, where both tests reject with probability 1 once imbalance is substantial enough. Yet, with more moderate levels of imbalance, the informativeness-weighted test correctly rejects as-if random with higher probability, as long as there is some non-trivial level of prognosis. We note that in the sampled natural

Table A3: Sufficient covariates: simulated rejection rates for each test

progX1	imbalX1	progX2	imbalX2	progX3	imbalX3	UW p	PW p	Hotelling p	Rsqr prog	Rsqr bal
0	0	0.25	0	0	0	0.001	0	0	1	0.002
0.2	0	0.25	0	0	0	0.001	0.001	0	1	0.002
0.4	0	0.25	0	0	0	0.002	0.002	0	1	0.002
0.6	0	0.25	0	0	0	0.002	0	0	1	0.002
0	0	0.25	0	0	0.05	0.021	0	0.013	1	0.004
0.2	0	0.25	0	0	0.05	0.016	0.001	0.005	1	0.004
0.4	0	0.25	0	0	0.05	0.017	0.001	0.008	1	0.004
0.6	0	0.25	0	0	0.05	0.013	0	0.005	1	0.004
0	0	0.25	0	0	0.1	0.163	0.002	0.253	1	0.012
0.2	0	0.25	0	0	0.1	0.133	0.001	0.246	1	0.012
0.4	0	0.25	0	0	0.1	0.143	0	0.223	1	0.012
0.6	0	0.25	0	0	0.1	0.134	0.002	0.243	1	0.011
0	0.05	0.25	0	0	0	0.012	0.001	0.003	1	0.004
0.2	0.05	0.25	0	0	0	0.016	0.022	0.007	1	0.004
0.4	0.05	0.25	0	0	0	0.015	0.056	0.006	1	0.004
0.6	0.05	0.25	0	0	0	0.016	0.063	0.013	1	0.004
0	0.05	0.25	0	0	0.05	0.15	0.001	0.054	1	0.007
0.2	0.05	0.25	0	0	0.05	0.13	0.018	0.041	1	0.007
0.4	0.05	0.25	0	0	0.05	0.129	0.039	0.04	1	0.007
0.6	0.05	0.25	0	0	0.05	0.147	0.066	0.036	1	0.007
0	0.05	0.25	0	0	0.1	0.529	0.001	0.394	1	0.015
0.2	0.05	0.25	0	0	0.1	0.478	0.024	0.4	1	0.015
0.4	0.05	0.25	0	0	0.1	0.481	0.055	0.39	1	0.015
0.6	0.05	0.25	0	0	0.1	0.506	0.057	0.413	1	0.015
0	0.1	0.25	0	0	0	0.134	0.001	0.261	1	0.012
0.2	0.1	0.25	0	0	0	0.138	0.181	0.262	1	0.012
0.4	0.1	0.25	0	0	0	0.136	0.46	0.256	1	0.0122
0.6	0.1	0.25	0	0	0	0.123	0.545	0.233	1	0.012
0	0.1	0.25	0	0	0.05	0.479	0.002	0.389	1	0.015
0.2	0.1	0.25	0	0	0.05	0.478	0.187	0.393	1	0.015
0.4	0.1	0.25	0	0	0.05	0.466	0.446	0.391	1	0.015
0.6	0.1	0.25	0	0	0.05	0.496	0.578	0.387	1	0.015
0	0.1	0.25	0	0	0.1	0.852	0.001	0.802	1	0.022
0.2	0.1	0.25	0	0	0.1	0.86	0.16	0.788	1	0.022
0.4	0.1	0.25	0	0	0.1	0.842	0.447	0.783	1	0.022
0.6	0.1	0.25	0	0	0.1	0.852	0.588	0.809	1	0.022

Note: Signal covariates are X1 and X2 and covariates included in the global tests are X1, X2, and X3. Covariate X2 is balanced in expectation and has a fixed parameter value for prognosis (0.25). Rejection rates are calculated over 1000 values of test statistic p-values. Further details of the data generating process can be found in Sections 7.1-7.2 of this appendix.

experimental studies in Figure 1 in the paper, the imbalance R^2s are mostly below 0.1. Thus, we would argue, this situation of relatively low imbalance is the one in which most need a powerful test of as-if random, and this is what the informativeness-weighted test delivers.

Conversely, when as-if random is true—and thus we do not wish to reject it—the Type I error rate of the unweighted test is higher. As the right panel shows, it rejects at least as often as the prognosis-weighted test when there is any non-zero level of prognosis; and it rejects more often when there is any imbalance on irrelevant (i.e., non-prognostic) covariates. (Note the absence of data points away from the axes reflects the structure of our simulation: with positive imbalance and positive prognosis, as-if random would be false).

In sum, we could think about these results in terms of three cases. First, when covariates are both highly prognostic and highly imbalanced, the weighted and unweighted tests reject with equal probability. Second, when there is high imbalance and low prognosis, the unweighted test may be more powerful when as-if random is false; but this runs the risk of spurious rejections when as-if random is true, as the right panel shows. Third and finally, however, when there is low imbalance and high prognosis, the informativeness-weighted test is both more powerful when as-if random is false and avoids spurious rejections when it is true.

The simulations confirm that compared to unweighted tests, the informativeness-weighted test can better detect the failure of as-if random while simultaneously limiting spurious rejections. In contrast, tests that do not take account of the relative prognosis of covariates are prone to reject as-if random when it is true or to fail to reject it when it is false, due to the balance or imbalance of irrelevant noise covariates. By projecting out irrelevant covariates and prioritizing prognostic ones, the informativeness-weighted test thus boosts both specificity and power.

7.3 Uninformative covariates

As noted in the paper, our simulations offer an important caveat. Consistent with our theoretical results, the quality of tests—including prognosis-weighted ones—depends on the joint prognosis of measured covariates.

In Figure A4, we consider simulations in which observed covariates are not sufficient, i.e., may be fully or partially uninformative about potential outcomes. Thus, as before, we measure a potentially prognostic covariate X_1 and the irrelevant noise covariate X_3. Now, however, a prognostic—and here, in contrast to the previous simulations, *imbalanced*—covariate X_2 is "omitted." Thus, potential outcomes are related to an unobserved as well as observed covariates, and the unobserved covariate is imbalanced as well as prognostic. In the four plots, we again vary the expected imbalance of the potentially prognostic variable $X1$ and the irrelevant noise covariate $X3$. Here, however—due to the prognosis and expected imbalance of the unmeasured X_2—as-if random is everywhere false. Table A2 gives a selection of the numerical results depicted graphically in Figure A3.

There are two key takeaways. First, failure to measure imbalanced, prognostic covariates can dramatically reduce the power of balance tests. In the top-left plot—when only the unobserved prognostic covariate is imbalanced—the weighted and unweighted tests both fail to detect the failure of as-if random. Here, the observed covariates X_1 and X_3 are both balanced, so failure of as-if random arises only due the imbalance of the prognostic X_2, which is omitted from the test. Both unweighted and prognosis-weighted tests reject at similarly low rates—producing a false negative rate of nearly 1.

Second and more reassuringly, however, the tests perform better when the signal covariates we do

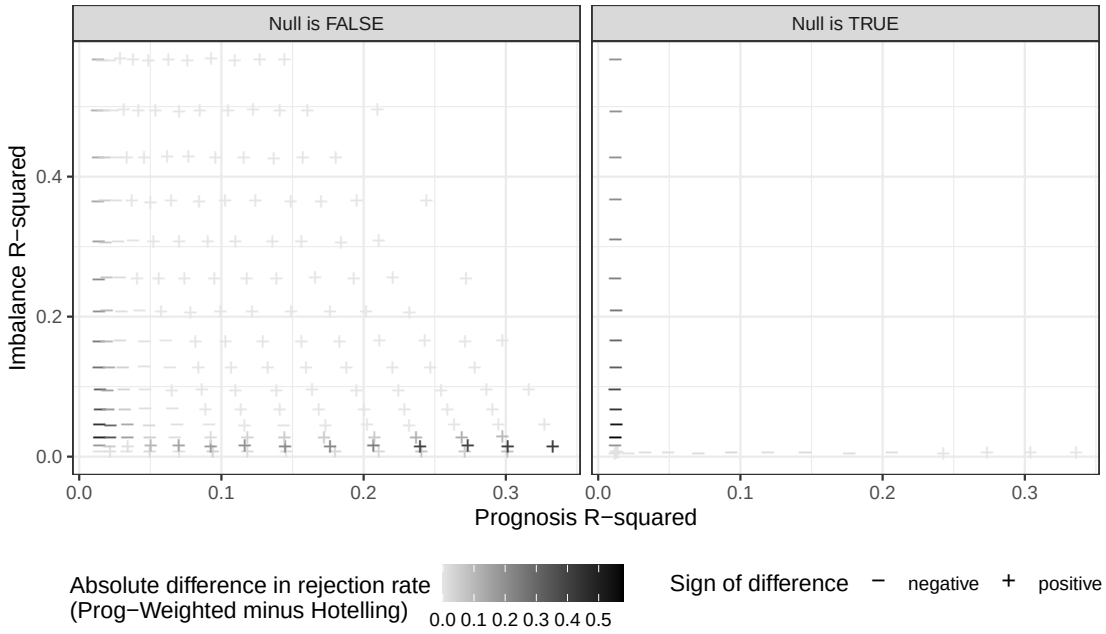


Figure A3: Broader Set of Simulations: Varying Imbalance and Prognosis

Difference in rejection rates of as-if random when it is false (left panel) and true (right panel), comparing the prognosis-weighted and unweighted (Hotelling's T^2) test. Positive (negative) signs indicate cases where the prognosis-weighted test has a higher (lower) rejection rate than Hotelling's T^2 test, with darker shading representing greater absolute difference.

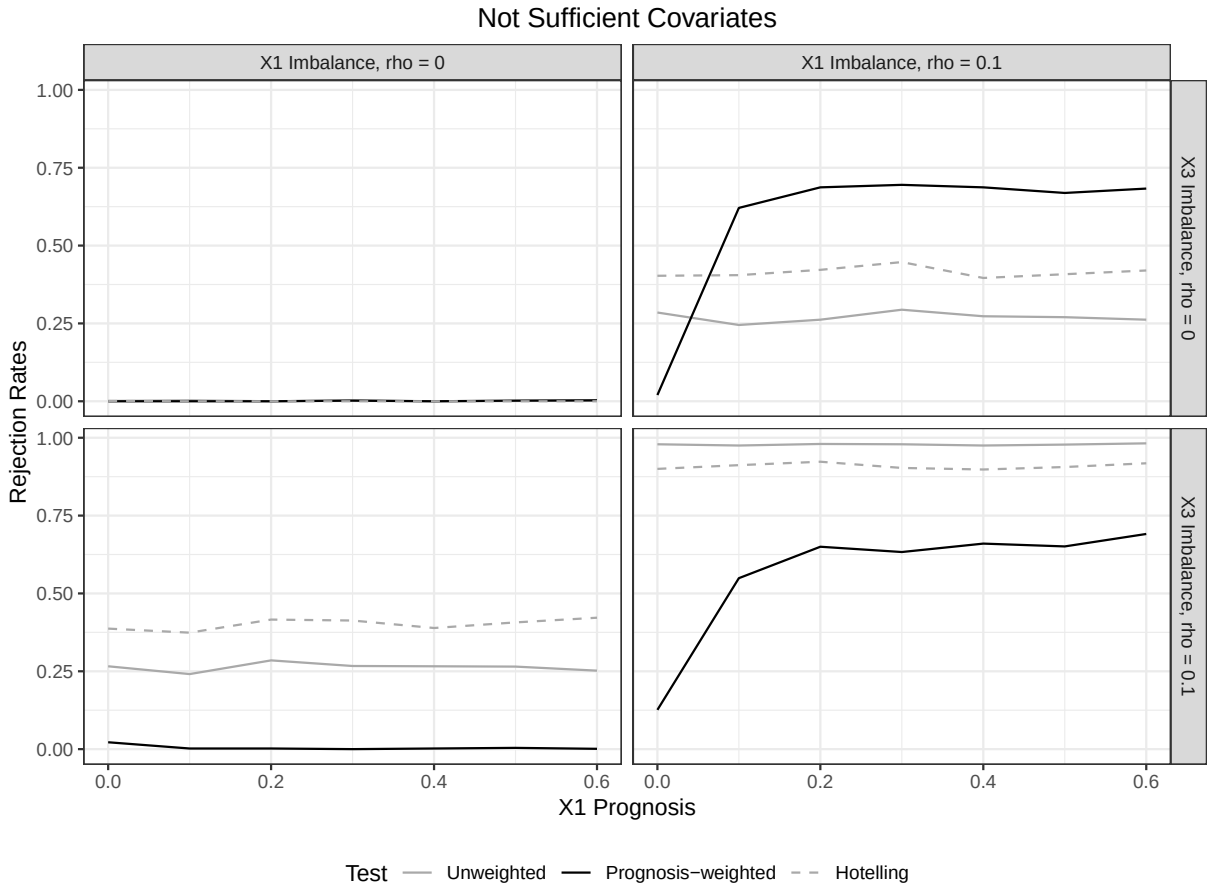


Figure A4: Covariates not sufficient: omitting a prognostic and imbalanced variable

Here, we consider tests in which a covariate X_2 that is prognostic ($\rho = 0.25$) and imbalanced ($\rho = 0.15$) in expectation is not measured. The tests make use only of a potentially prognostic covariate X_1 and non-prognostic noise covariate X_3. Here, as-if random is false in all cases. See subsection 7.3 of this Appendix for further details.

measure are increasingly prognostic. In the top-right plot, for example, as X_1 becomes more prognostic (both in absolute terms and relative to the fixed prognosis of X_2), the prognosis-weighted test becomes more powerful. Note that in the bottom two plots, where X_3 is imbalanced, the unweighted test rejects at rates even higher than the prognosis-weighted test. Given that as-if random is here false, this might suggest the desirability of a test that ignores prognosis, but this is misleading. The unweighted test rejects because it is wrongly sensitive to imbalance on the irrelevant covariate X_3. The difficulty is that in practice, we will not know if we are in a setting in which as-if random is false, as in Figure A4, or true—as it is for some parameter values in Figure A2, when the unweighted test erroneously rejects also due to its sensitivity to an irrelevant imbalanced covariate. The good news is that while the sensitivity of unweighted tests is invariant to prognosis, the weighted tests become increasingly powerful as we measure more relevant covariates. The more informative they are, the less likely the problem of unobserved prognostic covariates is to hinder the performance of the tests.

Table A4: Covariates not sufficient: Simulated rejection rates for each test

progX1	imbalX1	progX2	imbalX2	progX3	imbalX3	UW p	PW p	Hotelling p	Rsqr prog	Rsqr bal
0	0	0.25	0.15	0	0	0.001	0	0.001	0.004	0.001
0.2	0	0.25	0.15	0	0	0	0	0	0.398	0.001
0.4	0	0.25	0.15	0	0	0	0	0	0.723	0.001
0.6	0	0.25	0.15	0	0	0.002	0.003	0	0.855	0.001
0	0	0.25	0.15	0	0.05	0.024	0.001	0.015	0.004	0.004
0.2	0	0.25	0.15	0	0.05	0.029	0.001	0.016	0.395	0.004
0.4	0	0.25	0.15	0	0.05	0.035	0.002	0.022	0.724	0.004
0.6	0	0.25	0.15	0	0.05	0.019	0.002	0.014	0.855	0.004
0	0	0.25	0.15	0	0.1	0.266	0.022	0.387	0.004	0.011
0.2	0	0.25	0.15	0	0.1	0.285	0.002	0.416	0.397	0.012
0.4	0	0.25	0.15	0	0.1	0.266	0.002	0.389	0.724	0.011
0.6	0	0.25	0.15	0	0.1	0.252	0.001	0.422	0.855	0.012
0	0.05	0.25	0.15	0	0	0.017	0	0.02	0.004	0.004
0.2	0.05	0.25	0.15	0	0	0.022	0.067	0.013	0.393	0.004
0.4	0.05	0.25	0.15	0	0	0.024	0.101	0.018	0.721	0.004
0.6	0.05	0.25	0.15	0	0	0.023	0.082	0.024	0.853	0.004
0	0.05	0.25	0.15	0	0.05	0.252	0.005	0.087	0.004	0.006
0.2	0.05	0.25	0.15	0	0.05	0.268	0.081	0.102	0.394	0.006
0.4	0.05	0.25	0.15	0	0.05	0.288	0.094	0.095	0.722	0.006
0.6	0.05	0.25	0.15	0	0.05	0.283	0.09	0.107	0.854	0.006
0	0.05	0.25	0.15	0	0.1	0.741	0.031	0.579	0.004	0.014
0.2	0.05	0.25	0.15	0	0.1	0.76	0.073	0.6	0.392	0.014
0.4	0.05	0.25	0.15	0	0.1	0.79	0.077	0.623	0.721	0.014
0.6	0.05	0.25	0.15	0	0.1	0.737	0.075	0.561	0.855	0.014
0	0.1	0.25	0.15	0	0	0.285	0.02	0.403	0.004	0.012
0.2	0.1	0.25	0.15	0	0	0.262	0.687	0.422	0.386	0.012
0.4	0.1	0.25	0.15	0	0	0.273	0.687	0.396	0.718	0.011
0.6	0.1	0.25	0.15	0	0	0.262	0.683	0.42	0.851	0.012
0	0.1	0.25	0.15	0	0.05	0.745	0.054	0.58	0.005	0.014
0.2	0.1	0.25	0.15	0	0.05	0.739	0.654	0.563	0.386	0.014
0.4	0.1	0.25	0.15	0	0.05	0.756	0.669	0.581	0.719	0.014
0.6	0.1	0.25	0.15	0	0.05	0.755	0.678	0.581	0.852	0.014
0	0.1	0.25	0.15	0	0.1	0.979	0.126	0.9	0.004	0.021
0.2	0.1	0.25	0.15	0	0.1	0.98	0.65	0.923	0.387	0.022
0.4	0.1	0.25	0.15	0	0.1	0.975	0.66	0.898	0.718	0.022
0.6	0.1	0.25	0.15	0	0.1	0.982	0.691	0.918	0.852	0.022

Note: Signal covariates are X1 and X2 and covariates included in the global tests are X1 and X3. Omitted covariate X2 has fixed imbalance (0.15) and prognosis (0.25) parameter values. Rejection rates are calculated over 1000 values of test statistic p-values. Further details of the data generating process can be found in Section 7.3 of this appendix.

7.4 Performance of tests at threshold levels of prognosis

We can also look graphically at the sensitivity of tests across our full set of simulations, including those cases in which covariates are sufficient and those where they are not.

Figure A5 depicts the power of weighted tests as prognosis varies, with darker shading of points for higher power tests (and red points for those tests where power exceeds 80%). The null of as-if random is false in these simulations, so the shading indicates the probability that a false null is rejected. We include cases with sufficient covariates (left panel) and insufficient covariates (right panel). The horizontal axis measures joint prognosis of the covariates (the R^2 from the regression of control-group potential outcomes on covariates), while the vertical axis measures the realized imbalance in treatment assignment (the R^2 from the regression of treatment assignment on covariates). In these simulations, we also observe a possibly prognostic covariate X_1. In the "not sufficient" case (right panel), we add an observed noise covariate X_3 unrelated to potential outcomes and an unobserved signal covariate X_2 that is prognostic ($\rho = 0.25$) and imbalanced ($\rho = 0.15$) in expectation. Here, as-if random is false in all cases.

As expected, the power of the tests is excellent with sufficient covariates, reaching power in excess of 80% at low levels of imbalance (e.g. when the imbalance R^2 is less than 0.025). Perhaps more suprisingly, the tests also achieve similar levels of power with insufficient covariates as long as the prognosis R^2 is large enough. Even with a prognosis R^2 of 0.125, the tests achieve power in excess of 80% at when the imbalance R^2 is 0.025. Thus, even with low levels of expected imbalance, the test can detect failures of as-if random with high probability when covariates are sufficiently prognostic.

While the precise thresholds are surely a function of the data-generating process and parameters used in the simulations, it also appears that a conservative threshold would be a prognosis R^2 between 0.1 and 0.2 for adequately powered tests. Returning to Figure 1 in the paper, we also see that the top threshold lies at the upper bound of the lower prognosis tests but that many studies also do exceed the threshold.

7.5 Simulations under non-linearity

Our results so far have considered simulations in which potential outcomes are generated as linear functions of covariates. Because the conditional expectation function in the finite population is linear, a linear regression of control potential outcomes on covariates—in the control group sample—should well approximate the conditional expectation function, up to sampling error. While a simplifying assumption, this approach usefully allows us to compare prognosis-weighted to unweighted tests and to assess how the performance of tests varies as we modify the prognosis of covariates—both jointly and individually—as well as patterns of covariate imbalance.

In this section, we examine how our tests perform if the conditional expectation of potential outcomes, given covariates, is not linear. Specifically, we simulate three different scenarios to assess results under non-linearity in the finite-population regression function, i.e., in the relationship between covariates and potential outcomes. Thus, we consider settings in which the potential outcome regression function comprise

1. k-level polynomial terms (section 7.5.1);
2. covariate interactions (section 7.5.2); and
3. two 'difficult', highly non-linear functions, one based on a 'tree' specification and the other a 'sine' specification (section 7.5.3).

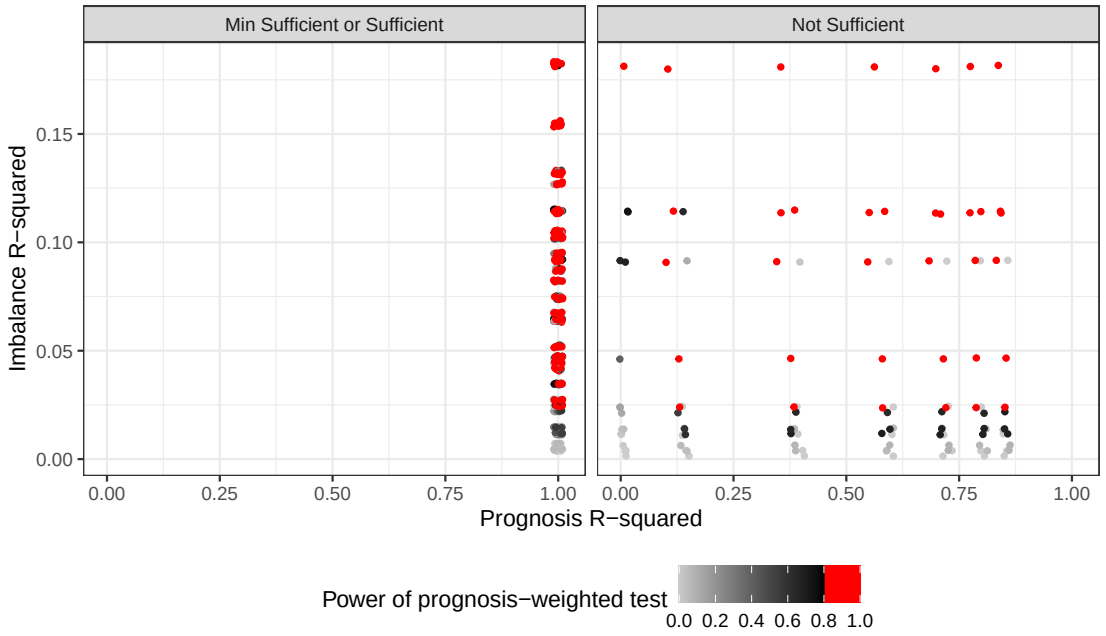


Figure A5: Power of weighted tests as prognosis varies

Here, we consider power in the full set of tests, including cases with sufficient covariates (left panel) and insufficient covariates (right panel). Shading indicates the probability that a false null is rejected. The horizontal axis measures joint prognosis of the covariates (the R^2 from the regression of control-group potential outcomes on covariates), while the vertical axis measures the realized imbalance in treatment assignment (the R^2 from the regression of treatment assignment on covariates). In these simulations, we observe a possibly prognostic covariate X_1. In the "not sufficient" case (right panel), we add an observed noise covariate X_3 unrelated to potential outcomes and an unobserved signal covariate X_2 that is prognostic ($\rho = 0.25$) and imbalanced ($\rho = 0.15$) in expectation. Here, as-if random is false in all cases.

Here, we expand the range of tests to include those based on linear models with polynomials of the covariates and/or interaction terms, as well as simple linear regressions; and machine learning techniques, specifically gradient boosted trees and random forests.

We also consider whether metrics such as the prognosis R^2 can allow us, by choosing the most accurate fitting procedure for $Y(0)$, potentially to improve performance by allowing the test to be based on the best-fitting procedure. Selection of the best-fitting model is automated in our `pwtest` package with the `contest` and `pick_winner` functions.

We emphasize that, in our setting, any discrepancies between (a) the conditional relationship of $Y(0)$ and covariates in the finite population and (b) the fitting procedure for $\widehat{Y(0)}$ are not issues primarily of biased inference. That is, our primary concern is not that $\mathrm{E}(\widehat{\beta}) \neq \beta$; but rather that improvements in the accuracy of fits of $\widehat{Y(0)}|X$ may improve test performance, in the sense of increasing power or specificity of the test of as-if random.

7.5.1 k-level polynomials in the potential outcomes model

In order to assess test performance under non-linearity, we first repeat the procedure described in Step 1 in subsection 7.1. That is, we generate a dataset of $N = 500$ observations. The dataset has a treatment assignment vector Z (with half the units assigned at random to treatment and half to control); potential outcomes $Y(1)$ and $Y(0)$; and covariates X_p, with $p = 1, 2$. Covariates are drawn from a multivariate normal distribution with mean 0 and standard deviation 1, and the elements of the variance-covariance matrix governing the variables are defined in such a way that the expected correlation between covariates and treatment assignment Z is determined by that covariate's imbalance parameter; the expected correlation between covariates is 0. As before, the average treatment effect is set to zero (i.e. $Y_i(0) = Y_1(1)$ for all i).

The key deviation from our previous approach is that we define potential outcomes under control as as a function of covariates such that

$$Y(0) = \boldsymbol{X_1}\theta + \beta X_2 \tag{52}$$

where $\boldsymbol{X_1}$ is an N by K matrix where each column is given by X_1^k, where $k = \{1, 2, ...K\}$, K being the highest polynomial term for X_1 in the regression. β and θ (K x 1 vector) are prognostic coefficients. When $k = 3$, for example, the data-generating process for potential outcomes is defined as:

$$Y(0) = \theta_1 X_1 + \theta_2 X_1^2 + \theta_3 X_1^3 + \beta X_2 \tag{53}$$

When $K = 1$, the simulation is mimics the approach we described above for Minimally Sufficient cases, so increasing K allows us to compare test performances under at least two different variations of non-linearity in potential outcome models.

We calculate the observed statistics δ_{UW}, δ_{PWLR}, and Hotelling's T^2 described in Steps 2 in Section 7.1 the same way as before, meaning we use the OLS regressions described in the main paper to obtain δ_{PWLR}. In addition, we calculate versions of δ_{PWLR} that allow for k-degree polynomials in the regression; call these statistics δ^k_{PWLR}, where $k = \{2, ...K\}$ so there are K total prognosis-weighted test statistics (including δ_{PWLR} where $k = 1$). We also calculate versions of δ_{PWLR} allowing covariate interactions, as described in the next sub-section (7.5.2), so that we can assess performance of the test when the non-linear test statistics do not match the underlying form of non-linearity in the data-generating process.

Finally, we also calculate the prognosis R^2 implied by the sample regression of $Y(0)$ on (a) covariates $\boldsymbol{X_1}$ and X_2, where we take k from 1 to K, which gives us K prognosis R^2s and (b) covariates X_1, X_2,

and their interaction (as in the next subsection). Thus, in each of the realizations of the data-generating process described in Step 1 above, we calculate the prognosis R^2 for a given k-level regression as defined by equation 52, yielding the R^2 *given the observed data*. The first R^2 measure therefore consists of the average of these values over 1000 iterations of the data-generating process described in Step 4 above.

We repeat Steps 1-4 in subsection 7.1 with different parameter values determining X_1 imbalance and prognosis. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation.

Results: polynomial simulations

Figure A6 shows simulation results, for a case where the data-generating process for potential outcomes includes the k-level polynomials. Here we consider the $k = 2$ case (i.e., linear and squared terms in the data-generating process), and we compare unweighted to prognosis-weighted tests, with and without the $k = 2$ polynomial.

Several results are noteworthy. First, in the left column, when there is zero expected imbalance on the potentially prognostic X_1 term, all of the tests control Type 1 error at similar rates.

Second, however, when the potentially prognostic variables X_1 and X_1^2 are imbalanced (second and third columns), we see greater divergence in the performance of different tests. Unweighted tests continue to exhibit the problems identified in subsection 7.2: because they do not prioritize prognostic covariates, they are prone to spurious rejections when as-if random is true but non-prognostic covariates are imbalanced. In addition, they generally have less power than the prognostic-weighted tests, particularly so as X_1 prognosis increases.

7.5.2 Covariate interactions in the potential outcomes model

We next probe the performance of our tests under an additional setting where linearity does not hold: in theory, nonlinearity may also arise from the covariates jointly (rather than independently) determining potential outcomes.

Under this approach, we define an alternative potential outcome model as follows:

$$Y(0) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) \tag{54}$$

The goal of this simulation is to assess and compare the performance of different tests (especially, the linear test without interactions and one that allows for interactions) in the presence of varied expected imbalances and prognosis of the main term X_1 and the interaction term or product term $X_1 * X_2$.

Generating independent imbalance on the main and interaction terms

It is nontrivial to design a simulation in which we can control the imbalance of the interaction term separately to the imbalance of the main terms.

In order to do this, we design a data-generating process based on Simpson's Paradox: the goal is that X1:X2 should be marginally balanced on X1 and X2, while nonetheless controlling the structure of the product so that the product is correlated with treatment.

Our goal is to independently control three distinct correlations: $\text{Cor}(Z, X_1)$, $\text{Cor}(Z, X_2)$, and $\text{Cor}(Z, X_1 * X_2)$.

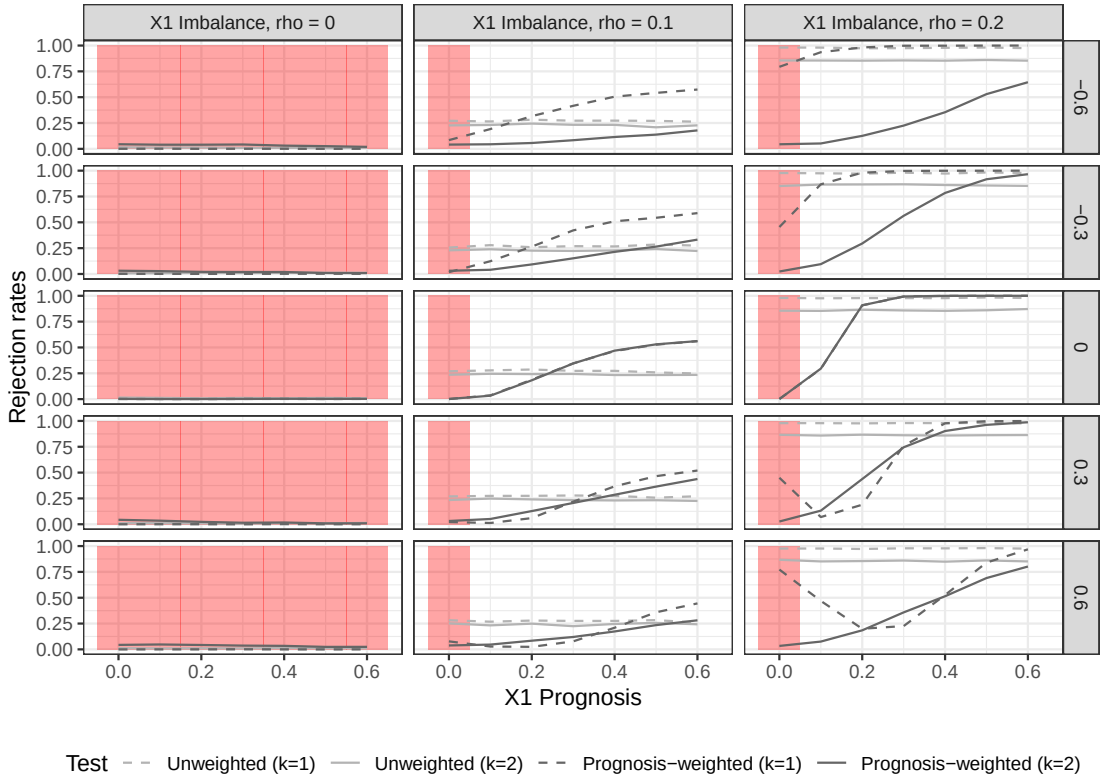


Figure A6: Rejection rates of tests with a second-degree polynomial in the data-generating process.

The x-axis of the Figure A6 shows values of X_1 prognosis and the y-axis shows rejection rates. The columns show results under different values of X1 imbalance, with rows showing different values of prognosis of X_1^2. The red bands highlight cases where the null hypothesis of as-if random holds.

The trick is first to notice that, if X1 and X2 are mean-centered, we can divide the covariate space $(X1, X2)$ into quadrants, and induce the desired correlation structure by manipulating the realized correlation between values in each quadrant and Z. We have the following three relationships:

$\text{Cor}(Z, X_1) \propto \mathbb{E}[Z|X1 > 0] - \mathbb{E}[Z|X1 < 0]$ – that is, $(Q1 + Q4) - (Q2 + Q3)$

$\text{Cor}(Z, X_2) \propto \mathbb{E}[Z|X2 > 0] - \mathbb{E}[Z|X2 < 0]$ – that is, $(Q1 + Q2) - (Q3 + Q4)$

$\text{Cor}(Z, X_1 \cdot X_2) \propto \mathbb{E}[Z|X1 \cdot X2 > 0] - \mathbb{E}[Z|X1 \cdot X2 \leq 0]$ – that is, $(Q1 + Q3) - (Q2 + Q4)$

Let p_1 denote the treatment probability for units in quadrant Q1 where $X_1 > 0$ and $X_2 > 0$, p_2 for units in quadrant Q2 where $X_1 \leq 0$ and $X_2 > 0$, p_3 for units in quadrant Q3 where $X_1 \leq 0$ and $X_2 \leq 0$, and p_4 for units in quadrant Q4 where $X_1 > 0$ and $X_2 \leq 0$.

We need to control the probabilities (p_1, p_2, p_3, p_4). To do so, we introduce probabilities $(\alpha_1, \alpha_2, \alpha_1 2)$, which parameterize the desired imbalance of X1, X2, and X1:X2 respectively. The goal is now to relate the quadrant probabilities to our parameters.

First, since correlation with X_1 depends on differences between quadrants where $X_1 > 0$ versus $X_1 \leq 0$, the parameter α_1 should appear with positive signs in Q1 and Q4 and negative signs in Q2 and Q3. Similarly, α_2 should have positive signs where $X_2 > 0$ (Q1 and Q2) and negative signs where $X_2 \leq 0$ (Q3 and Q4). Finally, α_{12} should have positive signs where $X_1 * X_2 > 0$ (Q1 and Q3, where both covariates have the same sign) and negative signs where $X_1 * X_2 < 0$ (Q2 and Q4, where the covariates have opposite signs).

Combining these sign patterns with a baseline probability $p_{\text{base}} = .05$ yields:

$$
\begin{aligned}
p_1 &= p_{\text{base}} + \alpha_1 + \alpha_2 + \alpha_{12} \\
p_2 &= p_{\text{base}} - \alpha_1 + \alpha_2 - \alpha_{12} \\
p_3 &= p_{\text{base}} - \alpha_1 - \alpha_2 + \alpha_{12} \\
p_4 &= p_{\text{base}} + \alpha_1 - \alpha_2 - \alpha_{12}
\end{aligned}
$$

provides the baseline treatment rate and α_1, α_2, and α_{12} are parameters controlling the desired correlations. This parameterization ensures that changing any single α parameter affects only one target correlation while leaving the others unchanged.

For the correlation with X_1, treatment assignment must differ systematically between units where $X_1 > 0$ versus $X_1 \leq 0$, because correlation fundamentally measures how two variables co-vary, and in this discrete setting, covariance reduces to the difference in conditional expectations across the relevant partition of the space. When we partition units by the sign of X_1, the correlation $\text{Cor}(Z, X_1)$ is directly proportional to $E[Z|X_1 > 0] - E[Z|X_1 \leq 0]$, since this difference captures how much treatment assignment systematically varies with the sign of X_1. Examining the treatment probability expressions, we see that α_1 appears with positive signs in Q1 and Q4 and negative signs in Q2 and Q3. Therefore, the expected treatment assignment conditional on $X_1 > 0$ is proportional to $(p_1 + p_4)/2 = p_{\text{base}} + \alpha_1$, while the expected assignment conditional on $X_1 \leq 0$ is proportional to $(p_2 + p_3)/2 = p_{\text{base}} - \alpha_1$. The difference between these expectations is $2\alpha_1$, establishing that α_1 directly controls the correlation between Z and X_1.

Similarly, the correlation with X_2 requires differences between units where $X_2 > 0$ versus $X_2 \leq 0$, creating a horizontal partition that groups Q1 and Q2 against Q3 and Q4. In the probability expressions, α_2 appears with positive signs in Q1 and Q2 and negative signs in Q3 and Q4, yielding expected assignments

of $p_{\text{base}} + \alpha_2$ for $X_2 > 0$ and $p_{\text{base}} - \alpha_2$ for $X_2 \leq 0$. Again, the difference is $2\alpha_2$, so α_2 controls the correlation with X_2.

Since $X_1 * X_2 > 0$ when X_1 and X_2 have the same sign, this occurs in quadrants Q1 (both positive) and Q3 (both negative). Conversely, $X_1 * X_2 < 0$ when X_1 and X_2 have opposite signs, which occurs in Q2 and Q4. This creates a diagonal partition of the space. The parameter α_{12} appears with positive signs in Q1 and Q3 and negative signs in Q2 and Q4, yielding expected assignments of $p_{\text{base}} + \alpha_{12}$ when $X_1 * X_2 > 0$ and $p_{\text{base}} - \alpha_{12}$ when $X_1 * X_2 < 0$. The difference is $2\alpha_{12}$, establishing control over the interaction correlation.

The independence of these three correlations follows from the orthogonality of the underlying sign patterns. If we represent each α parameter's influence as a vector over the four quadrants, we have, for quadrants $(Q1, Q2, Q3, Q4)$, $\alpha_1 \sim (+1, -1, -1, +1)$ $\alpha_2 \sim (+1, +1, -1, -1)$, and $\alpha_{12} \sim (+1, -1, +1, -1)$. These vectors are orthogonal in the sense that their dot products are zero, which means that changing one parameter does not affect the correlations controlled by the others. This orthogonality is what allows the simulation to independently manipulate the balance of main effects versus interaction effects.

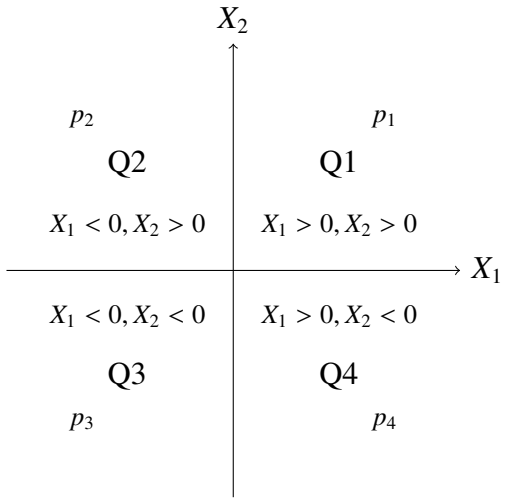


Figure A7: Quadrant-based paramaterization of covariate space to induce independent control of imbalance in $X1$, $X2$, and $X1 \cdot X2$. By dividing the space $(X1, X2)$ into quadrants, and paramaterizing the probability of treatment in each quadrant, we can design a simulation in which $Cor(Z, X1)$, $Cor(Z, X2)$ and $Cor(Z, X1 \cdot X2)$ are parameterized separately.

Steps in the interaction simulations

We repeat Steps 1-4 in Section 7.1 with different parameter values determining the imbalance and prognosis of X_1 and the interaction term $X_1 * X_2$.

We consider four cases of prognosis: one in which standardized coefficients on X_1 and $X_1 * X_2$ are both set to 0.25 in the data-generating process; one in which X_1 prognosis is 0.25 and $X_1 * X_2$ is 0.5; one in which X_1 prognosis is 0.5 and $X_1 * X_2$ is 0.25; and one in which X_1 and $X_1 * X_2$ are both set to 0.5.

For each of these cases, following the procedure described above, we vary the expected imbalance of X_1 and $X_1 * X_2$ factorially, in the sequence $\rho = 0.0, 0.1, 0.2$. This gives $3 \times 3 = 9$ combinations of X_1 and $X_1 * X_2$ imbalance for each of the four cases of prognosis. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation.

Finally, for each of these combinations of covariate prognosis and imbalance, we plot the rejection rates for the linear test (excluding interactions) and the test based on the interactive regression of $Y(0)$ on X_1, X_2, and their product. We take the prognosis R^2 in each regression so that we can assess the performance of the test based on the best fitting regression to the tests based on worse fitting regressions.

Interaction simulation results

Figures A8-A11 depict the results. Several conclusions are noteworthy.

First, in the top-left facet of all four prognosis cases in Figures A8-A11, the linear and interaction models appropriately control Type I error. Here, both the main X_1 and X_2 terms as well as the interaction term $X_1 * X_2$ are balanced in expectation, so as-if random hold.

Second, in the other facets where as-if random does not hold (because of expected imbalance on X_1 or $X_1 * X_2$ or both) we see that the rejection rate of prognosis-weighted test is generally increasing in the empirical prognosis R^2 of the model with an interaction. The interactive test generally outperforms the linear test without an interaction except when the product $X_1 * X_2$ is balanced in expectation (top row of Figures A8-A11).

Third, comparing the first, second, and third column in each figure, we see that the better performance of the interactive test, relative to the linear test, is most pronounced in those cases where the main X_1 term is balanced but the interactive term $X_1 * X_2$ is imbalanced (middle and bottom facets of first column). As the imbalance of the linear term grows, the performance gap shrinks. As noted, the linear test outperforms the interactive test when the interactive term is balanced (middle and right facets of the top row). In the second and third rows, where there is imbalance on both the main and interactive term, the interactive test outperforms the linear test, even when imbalance is greater on the linear term (e.g., the bottom facet of the middle column). However, when imbalance is substantial on both terms (e.g., the middle and bottom facets of the final column), the difference in the performance of the linear and interactive tests is minor, even when the product term is more imbalanced than the main term.

Finally, comparing across Figures A8-A11, we can also see that the gap between the performance of the interactive and linear tests is greatest when the product term is relatively prognostic, compared to the main term: compare for instance Figure A9, where the prognosis of the interactive term is 0.5 compared to 0.25 for the main term, and Figure A10, where the prognosis of the interactive term is instead 0.25 compared to 0.5 for the main term. In the former, the gap in rejection rates is substantial when the product term is also more imbalanced than the main term (middle and bottom facets of the middle column in Figure A9). In contrast, when the main term is relatively prognostic compared to the interactive term, the performance of the tests is nearly indistinguishable even when the interactive term is more highly imbalanced.

One other feature of the simulation results that may be noteworthy is the excellent performance of the tests, even at relatively low prognosis R^2s. For example, when the main and interactive terms are both prognostic and imbalanced (e.g. the middle and bottom facet of the final column in each figure), both tests reach rejection rates near 1 at empirical R^2s between 0.02 and 0.06. While this may depend on specifics of the simulation and data-generating process, it broadly reinforces the conclusion discussed in section 7.4, in which we saw that the tests could reach adequate performance at moderate levels of prognosis. (In the simulations in section 7.4, power of 80% was reaching with prognosis R^2 between 0.1 and 0.2).

In sum, the conclusions are thus broadly similar to those for the polynomial case in subsection 7.5.1. The linear test often does quite well, relative to the interactive test, as long as the main term is both imbalanced in expectation and prognostic. In contrast, in those cases where there is greater imbalance on

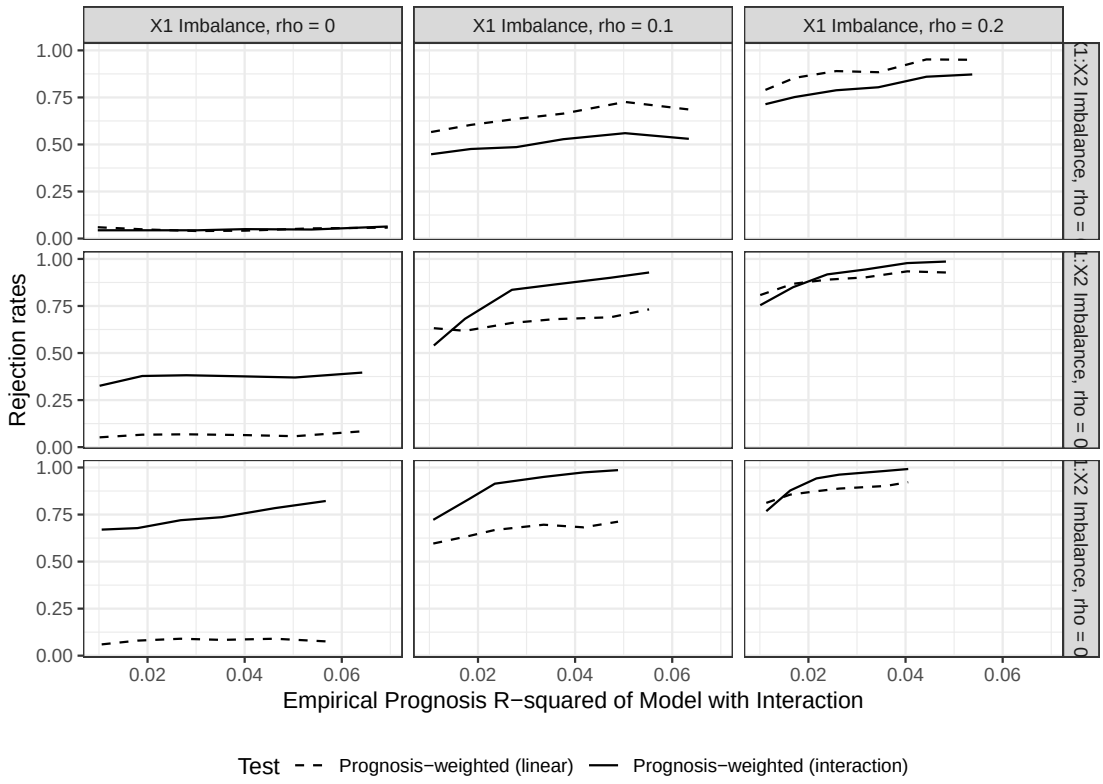


Figure A8: **Rejection rates of prognosis weighted test in the presence of a interaction term between X_1 and X_2 in the data generating process (equal X_1 and $X_1 * X_2$ prognosis of 0.25).** The horizontal axis shows the empirical prognosis R^2 of the interaction model, and the left vertical axis rejection rates, as defined above. The row headers show the expected imbalance of X_1 (i.e. the main linear term in the d.g.p.), in a sequence from 0 to 0.1 to 0.2, and the column headers (on the right side of the plot's vertical axis) show the expected imbalance $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.), in the same sequence. Thus, the rows plot results under different values of X1 imbalance, while columns plot results under different values of $X_1 * X_2$ imbalance. As-if random holds in the upper-left facet, where both X_1 and the product of X_1 and X_2 are balanced in expectation. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation. In this simulation, the prognosis of $X1$ (the coefficient on the linear X_1 term in the d.g.p.) and the prognosis of $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.) are both fixed at 0.25.

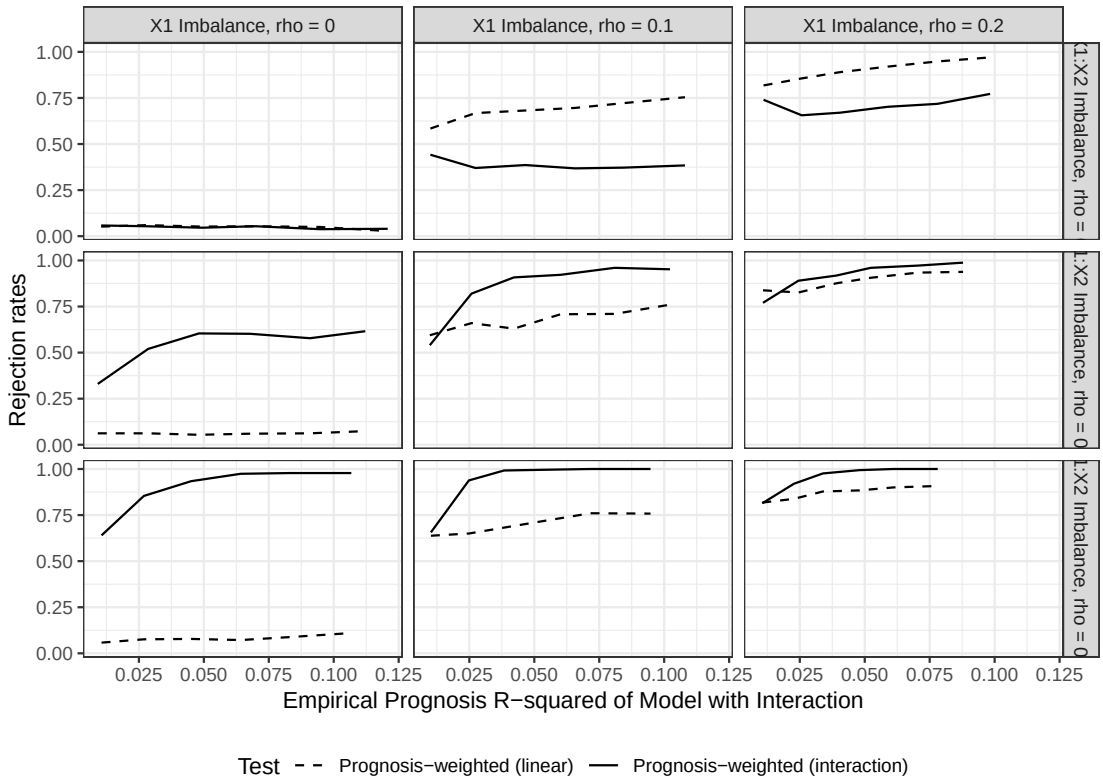


Figure A9: **Rejection rates of prognosis weighted test in the presence of a interaction term between X_1 and X_2 in the data generating process (X_1 prognosis of 0.25 and and $X_1 * X_2$ prognosis of 0.5).** The horizontal axis shows the empirical prognosis R^2 of the interaction model, and the left vertical axis rejection rates, as defined above. The row headers show the expected imbalance of X_1 (i.e. the main linear term in the d.g.p.), in a sequence from 0 to 0.1 to 0.2, and the column headers (on the right side of the plot's vertical axis) show the expected imbalance $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.), in the same sequence. Thus, the rows plot results under different values of X1 imbalance, while columns plot results under different values of $X_1 * X_2$ imbalance. As-if random holds in the upper-left facet, where both X_1 and the product of X_1 and X_2 are balanced in expectation. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation. In this simulation, the prognosis of $X1$ (the coefficient on the linear X_1 term in the d.g.p.) is 0.25 and the prognosis of $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.) is at 0.5.

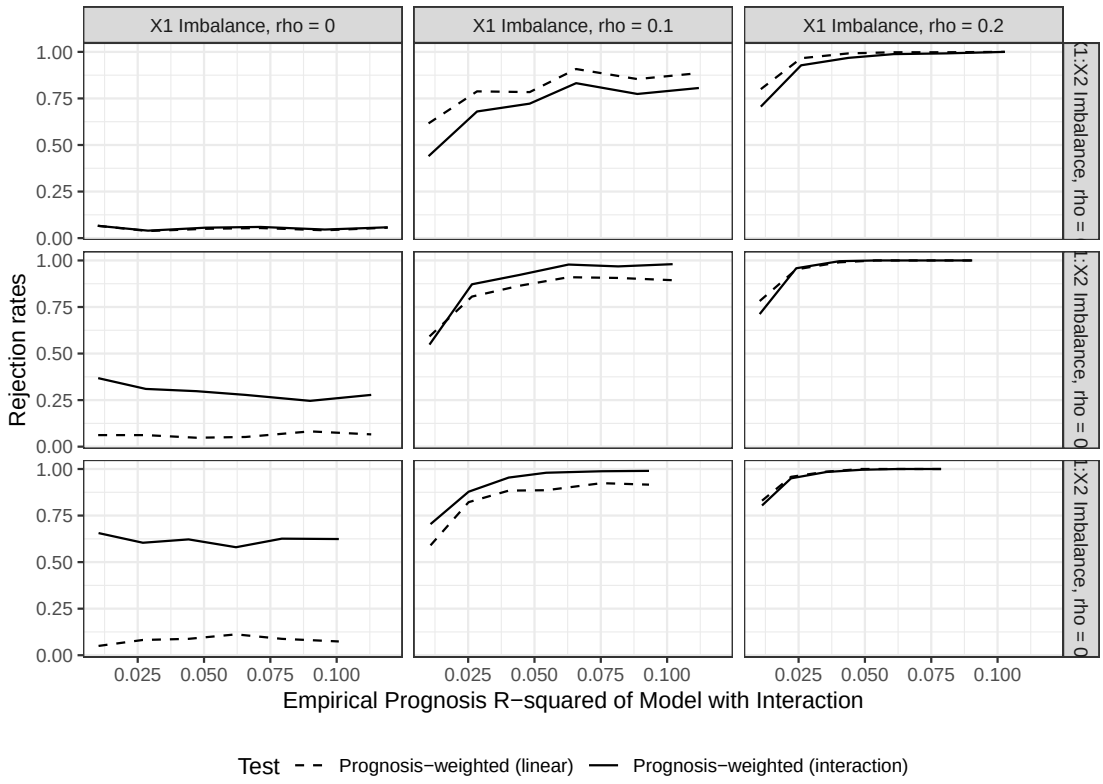


Figure A10: **Rejection rates of prognosis weighted test in the presence of a interaction term between X_1 and X_2 in the data generating process (X_1 prognosis of 0.5 and and $X_1 * X_2$ prognosis of 0.25).** The horizontal axis shows the empirical prognosis R^2 of the interaction model, and the left vertical axis rejection rates, as defined above. The row headers show the expected imbalance of X_1 (i.e. the main linear term in the d.g.p.), in a sequence from 0 to 0.1 to 0.2, and the column headers (on the right side of the plot's vertical axis) show the expected imbalance $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.), in the same sequence. Thus, the rows plot results under different values of X1 imbalance, while columns plot results under different values of $X_1 * X_2$ imbalance. As-if random holds in the upper-left facet, where both X_1 and the product of X_1 and X_2 are balanced in expectation. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation. In this simulation, the prognosis of $X1$ (the coefficient on the linear X_1 term in the d.g.p.) is 0.5 and the prognosis of $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.) is at 0.25.

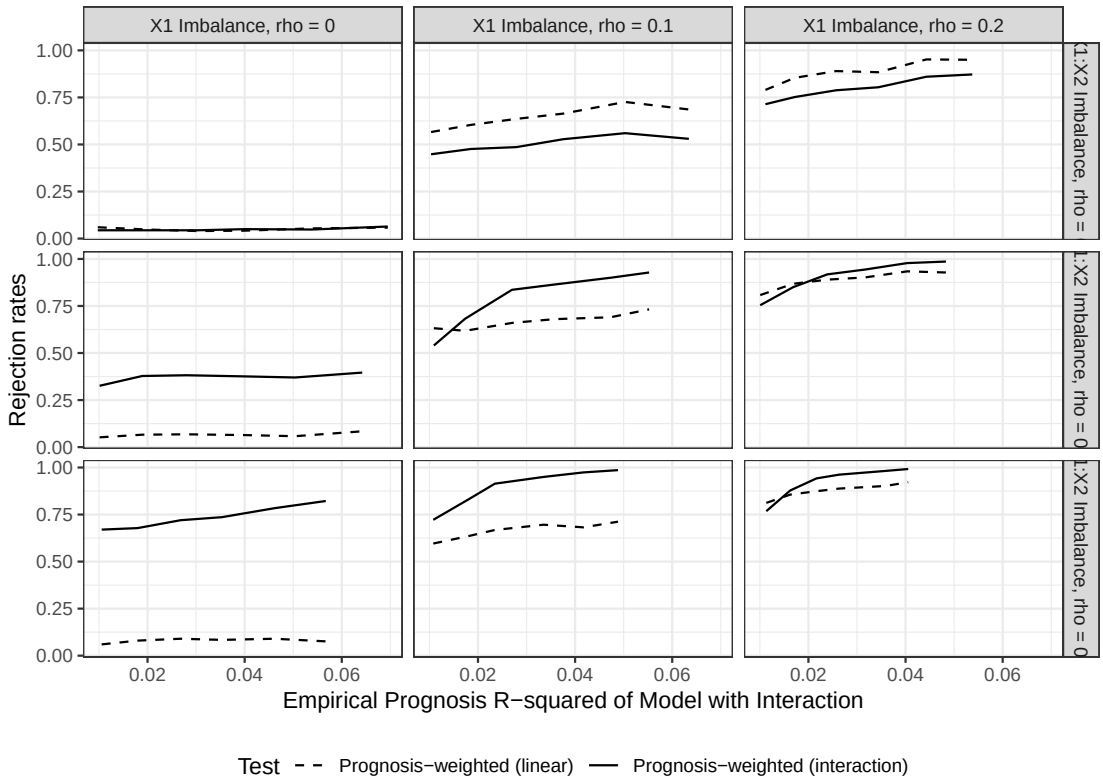


Figure A11: **Rejection rates of prognosis weighted test in the presence of a interaction term between X_1 and X_2 in the data generating process (equal X_1 and $X_1 * X_2$ prognosis of 0.5).** The horizontal axis shows the empirical prognosis R^2 of the interaction model, and the left vertical axis rejection rates, as defined above. The row headers show the expected imbalance of X_1 (i.e. the main linear term in the d.g.p.), in a sequence from 0 to 0.1 to 0.2, and the column headers (on the right side of the plot's vertical axis) show the expected imbalance $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.), in the same sequence. Thus, the rows plot results under different values of X1 imbalance, while columns plot results under different values of $X_1 * X_2$ imbalance. As-if random holds in the upper-left facet, where both X_1 and the product of X_1 and X_2 are balanced in expectation. The variable X_2 is set to have a fixed prognosis of 0.25 ($\beta = 0.25$) and is always balanced in expectation. In this simulation, the prognosis of $X1$ (the coefficient on the linear X_1 term in the d.g.p.) and the prognosis of $X_1 * X_2$ (the coefficient on the interaction term in the d.g.p.) are both fixed at 0.5.

the interactive term alone, and that term is more prognostic in the data-generating process for $Y(0)$ than the main term, we see more divergence in performance. In many cases, it may therefore be adequate to use the simple and interpretable linear test; yet analysts should consider the substantive domain under consideration and be attentive to the possibility of non-linear imbalances on prognostic variables. The issue of non-linear prognostic imbalance has not received attention in previous work on covariate balance testing, to our knowledge, so that is a contribution of our work.

7.5.3 Complex DGPs

Additionally, we evaluate the testing methods on two 'difficult' data-generating processes with highly nonlinear structure.

The simulation implements two distinct data generating processes for potential outcomes under control. The "tree" specification creates regime-dependent relationships based on the sign of the interaction term:

$$Y_0 = \begin{cases} \beta_1 X_1^2 + \beta_2 X_2 & \text{if } X_1 * X_2 > 0 \\ \beta_3 X_1 + \beta_4 X_2^2 & \text{if } X_1 * X_2 \leq 0 \end{cases}$$

This specification tests the method's performance when the functional form switches discretely based on the interaction term, creating fundamentally different covariate-outcome relationships across regions of the space. The "sine" specification incorporates high-frequency nonlinearities:

$$Y_0 = \beta_1 X_1 + \beta_2 \sin(5X_1) + \beta_3 X_2^2 + \beta_4 X_1 * X_2$$

This formulation challenges linear prognosis models with oscillatory components that standard polynomial approximations cannot capture well.

Imbalance is implemented by constructing a continuous score Z_{score} as:

$$Z_{\text{score}} = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 (X_1^2 - 1) + \alpha_4 (X_2^2 - 1) + \alpha_5 X_1 * X_2 + \epsilon$$

where $\epsilon \sim N(0, 1)$ provides random variation and the α parameters correspond to the desired correlations with X_1, X_2, X_1^2, X_2^2, and $X_1 * X_2$ respectively. The quadratic terms are centered by subtracting 1 to reduce their mechanical correlation with the linear terms, since $E[X_i^2] = 1$ when $X_i \sim N(0, 1)$. Binary treatment assignment follows from $Z = \mathbf{1}(Z_{\text{score}} > \text{median}(Z_{\text{score}}))$, ensuring exactly half the population receives treatment while preserving the correlation structure embedded in the scoring function.

Each of the simulations above follows the same orthogonalization procedure to achieve the target R^2 (level of prognosis). The signal component is standardized, orthogonal noise is generated through regression residuals, and the final potential outcomes are constructed as

$$Y_0 = \lambda Y_{0,\text{signal}} + \sqrt{1 - \lambda^2} Y_{0,\text{noise}}$$

where $\lambda = \sqrt{R^2_{\text{target}}}$.

Results with complex nonlinear data-generating process

Figures A12 and A13 depict the results. As in our previous simulations, the prognosis-weighted based on "saturated" linear regressions—i.e., those with expanded polynomial bases or covariate interactions—offer improvements in power both over simple linear methods and, perhaps surprisingly, the machine

learning methods and the best-fitting method chosen in each run of the simulation. This is likely due to greater stability: the machine-learning and best-fitting methods chosen in the control group may lead to a form of overfitting, limiting power when results are extrapolated to the treatment group. This should be studied in further simulations with greater numbers of covariates. However, our results here ven with the complex, 'difficult' data-generating processes, the expanded linear model with polynomial bases and covariate interactions has the greatest power. Along with the greater simplicity and interpretability of the prognosis weights in the linear methods, this lead to a general preference tests based on (saturated) linear fits.

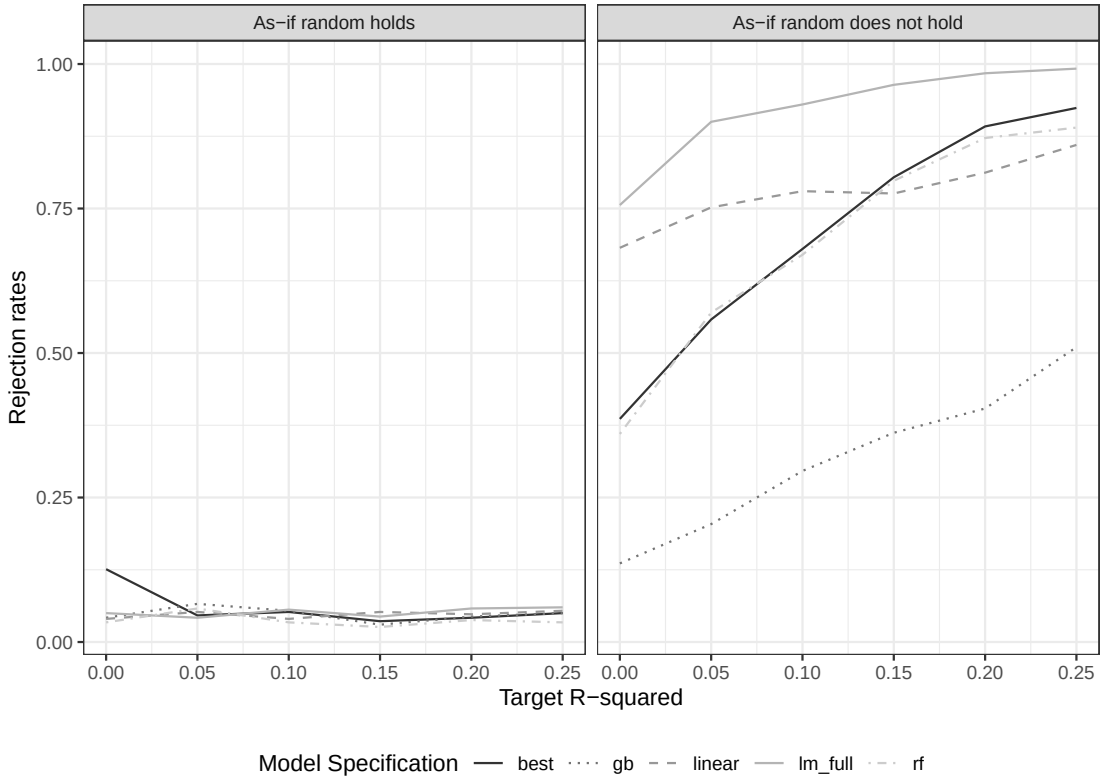


Figure A12: **Rejection rates of prognosis weighted test in the presence of the complex DGP using the "tree" specification described in Section 7.5.3.** All covariates are either balanced in expectation across treatment and control (left panel) or equally imbalanced in expectation (right panel). The target R^2 parameter in the x-axis takes the values in $\{0, 0.05, 0.1, 0.15, 0.20, 0.25\}$. Model specifications for fitting $\widehat{Y}$ include a tuned gradient-boosting model ("gb"), a simple linear regression using X_1 and X_2 ("linear"), a regression also including second degree polynomials and interactions for X_1 and X_2 ("lm_full"), and a tuned random forest model. We also display rejection rates for the "best" model (from all listed) for each instance of the data-generating process based on R^2 values of the prognosis regression fit.

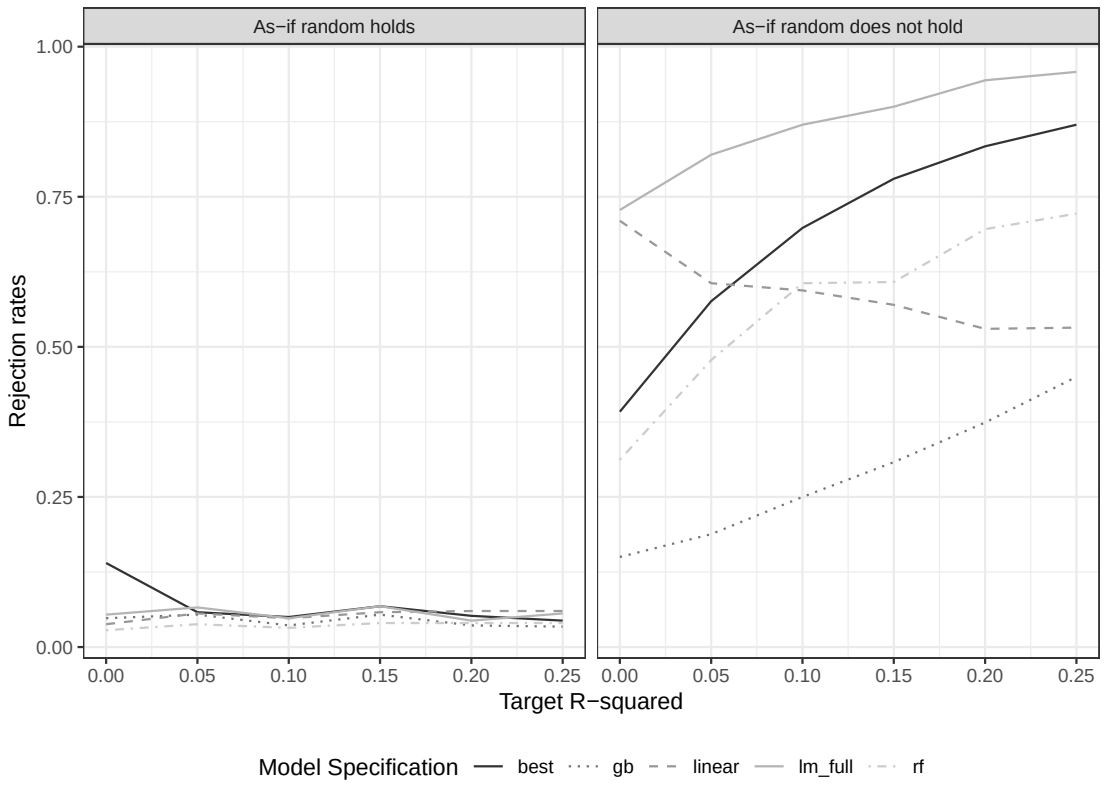


Figure A13: **Rejection rates of prognosis weighted test in the presence of the complex DGP using the "sine" specification described in Section 7.5.3.** All covariates are either balanced in expectation across treatment and control (left panel) or equally imbalanced in expectation (right panel). The target R^2 parameter in the x-axis takes the values in $\{0, 0.05, 0.1, 0.15, 0.20, 0.25\}$. Model specifications for fitting $\widehat{Y}$ include a tuned gradient-boosting model ("gb"), a simple linear regression using X_1 and X_2 ("linear"), a regression also including second degree polynomials and interactions for X_1 and X_2 ("lm_full"), and a tuned random forest model. We also display rejection rates for the "best" model (from all listed) for each instance of the data-generating process based on R^2 values of the prognosis regression fit.

7.6 Simulations: main takeaways

Overall, the lessons of the simulations for researchers are straightforward.

First, the prognosis of covariates is fundamental for the overall performance of balance tests. While there is no absolute threshold of required prognosis—as this will depend on the data-generating process—our simulations suggest improvements in power once the prognosis R^2 reaches a fairly minimimal 0.125. In general, a prognosis in excess of 0.2 would surely be desirable, but the main rule is that analysts should endeavor to measure the most prognostic covariates possible. This will maximize the prognosis R^2, which should be reported.

Second, given a particular set of measured (and ideally jointly prognostic) covariates, we can most effectively limit both false positives and false negatives simultaneously by prioritizing the most informative individual covariates—as in our prognosis-weighted test. Our test procedure essentially attempts to create a minimally sufficent set by prioritizing the most prognostic covariates for testing. As covariates overall become more prognostic, projecting out irrelevant covariates and focusing on the relatively prognostic ones can avoid both false positives and false negatives.

Thus, our results illustrate how prognosis weighting can reduce both false negatives and false positives. The performance of the tests depends on the overall prognosis of measured covariates. In contrast, unweighted tests that do not use information on covariate prognosis sacrifice power and/or specificity.

8 Software implementation: `R` package `pwtest`

8.1 Overview of `pwtest`

The package `pwtest` implements the prognosis-weighted test we propose and offers a visualization tool for diagnosing covariate-by-covariate balance and prognosis. Users can easily extract the global p-values from our test for easy reporting.

We offer three main functions:

- `pwtest()` produces unweighted (optional) and prognosis-weighted statistics with standard errors and p-values for the *test of as-if random*.
- `pwtest_rdd()` produces unweighted (optional) and prognosis-weighted statistics with standard errors and p-values for the *test of continuity* (RD designs).
- `prog_bal()` generates a plot of standardized covariate difference-in-means in the y-axis and prognosis weights as standardized coefficients from regression of outcome on control units. The plot offers a visual diagnostic for covariate-by-covariate standardized difference in means and prognosis weights from standardized coefficients of prognostic regression of $Y^C(0)$ on set of covariates. The black dots show covariates with significant p-values ($\alpha = 0.05$) two-tailed t-tests of difference in means. The red triangle indicates the value of the R-squared from the prognosis regression on the x-axis and the balance regression on the y-axis.

8.2 Treatment of missing data

The software implements our baseline prognosis-weighted test of as-if random in such a way as to preserve as much covariate data as possible. This is important because in applications, covariate data is often

missing; moreover, the missingness is uneven across different covariates. Standard covariate-by-covariate balance tests therefore may have different effective sample (study group) sizes for different covariates. Taking differences of means separately for each covariate preserves these differences, with variance calculations and p-values for hypothesis tests reflecting particular effective sample size for each covariate.

We replicate this approach in our software implementation. The test statistic δ_{PWLR}, for example, is calculated as the vector product of prognosis weights, estimated in the control group, and the vector containing differences of means for each covariate. Thus, the differences of means across treatment and control is calculated separately for each covariate, using all available data (as long as the treatment indicator is not missing for the covariate, so that control and treatment observations can be distinguished). This applies also to polynomial and interaction versions of the regression-based test.

This is an important point to underscore because we found that in applications, results can be highly sensitive the treatment of missing data (for instance, using listwise deletion so that only cases with observations for all covariates are included). See, for instance, discussion of analysis of Caughey and Sekhon (2011) in section 5 of the paper.

For tests of the continuity assumption in RD designs as well as the machine learning approaches, the approach differs somewhat. For instance, in tests using δ_{PW}^{RD}, we use `rdrobust` for fitting the prognosis-weighted intercepts and thus inherit the listwise-deletion defaults.

8.3 Installation instructions

```
# install development version
devtools::install_github("clarabicalho/pwtest")
```

8.4 Usage example

Below we show code that can be used to install the package `pwtest` to implement our prognosis-weighted tests using the `pwtest` function.

We offer an example of the code syntax using open-source data available from Caughey and Sekhon (2011). The code produces the results for the Caughey and Sekhon study reported in Figure 1 in the paper.

The data were downloaded in .Rdata format from the Harvard Dataverse at
https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/8EYYA2/DRWB57.

```
library(pwtest)

# load reproduction data
load(file = "RDReplication.RData")

# filter as per RDrdobReplication.do
cs_asif <- filter(x, Use == 1 & abs(DifDPct) < .5 & !is.na(DifDPct))
cs_rd <- filter(x, Use == 1 & !is.na(DifDPct))

# as-if random test
delta_np <- pwtest(data = cs_asif,
                   covariates = c("DWinPrv", "DPctPrv", "DifDPPrv",
```

```
                        "IncDWNOM1", "ElcSwing", "DemInc", "NonDInc",
                        "PrvTrmsD", "PrvTrmsO", "RExpAdv", "DExpAdv",
                        "SoSDem", "GovDem", "VtTotPct"),
                      treatment = "DemWin",
                      outcome = "DPctNxt",
                      nsims = 500)

# as-if random test results
delta_np$estimates

# testing continuity of potential outcome in RD design
delta_rd <- pwtest_rdd(
    data = cs_rd,
    covariates = c("DWinPrv", "DPctPrv", "DifDPPrv",
                   "IncDWNOM1", "ElcSwing", "DemInc", "NonDInc",
                   "PrvTrmsD", "PrvTrmsO", "RExpAdv", "DExpAdv",
                   "SoSDem", "GovDem", "VtTotPct"),
    treatment = "DemWin",
    running_var = "DifDPct",
    outcome = "DPctNxt",
    nsims = 500,
    se_type = "bootstrap"
)

# continuity test results
delta_rd$estimates

# example graph
plot_pbal(delta_np, label_option = "minmax",
          show_color_legend = TRUE)
```

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arboretti, R., Carrozzo, E., Pesarin, F., and Salmaso, L. (2018). Testing for equivalence: an intersection-union permutation solution. *arXiv:1802.01877v1 [stat.AP] 6 Feb 2018*.

Bueno, N., Dunning, T., and Tuñón, G. (2014). Design-based analysis of regression discontinuities: Evidence from an experimental benchmark. Social Science Research Network.

Bueno, N. S. and Tuñón, G. (Spring 2015). Graphical presentation of regression discontinuity results. *The Political Methodologist (Newsletter of the Political Methodology Section of the American Political Science Association)*, 22(2).

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. *R Journal*, 7(1):38–51.

Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (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, M. D., Idrobo, N., and Titiunik, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Caughey, D., Dafoe, A., and Seawright, J. (2017). Nonparametric combination (npc): A framework for testing elaborate theories. *The Journal of Politics*, 79(2):688–701.

Caughey, D. and Sekhon, J. S. (2011). Elections and the regression discontinuity design: Lessons from close u.s. house races, 1942-2008. *Political Analysis*, 19(4):385–408.

Cochran, W. 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.

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

Dunning, T. (2008). Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly*, 61(2):282–293.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Strategies for Social Inquiry. Cambridge University Press.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., and Snyder, J. M. J. (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.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Freedman, D. A. (1999). From association to causation: some remarks on the history of statistics. *Statistical Science*, 14(3):243 – 258.

Freedman, D. A., Pisani, R., and Purves, R. (2007). *Statistics*. W.W.Norton, fourth edition.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Co.

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

Hartman, E. and Hidalgo, F. D. (2018). An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79:933–959.

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

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Liao, L. D., Zhu, Y., Ngo, A. L., Chehab, R. F., and Pimentel, S. D. (2023). Using joint variable importance plots to prioritize variables in assessing the impact of glyburide on adverse birth outcomes.

Lin, W., Dudoit, S., Nolan, D., and Speed, T. P. (2023). From urn models to box models: Explaining neyman's (1923) "minor miracle". Department of Statistics, University of California, Berkeley and Wharton School, University of Pennsylvania.

Liu, K. and Ruan, F. (2020). A self-penalizing objective function for scalable interaction detection.

Neyman, J. S., Dabrowska, D. M., and Speed, T. P. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (Translated 1990). *Statistical Science*, 5(4):465 – 472.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rubin, D. B. (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.

Samii, C. and Aronow, P. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters - STAT PROBAB LETT*, 82.

Sekhon, J. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. *Advances in Econometrics*, 38:1–28.

Stuart, E., Lee, B., and Leacy, F. (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.

Wang, Y. and Wang, L. (2020). Causal inference in degenerate systems: An impossibility result. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3383–3392. PMLR.