The power of prognosis: Informativeness-weighted covariate balance tests

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

June 29, 2023

Abstract

Scholars use covariate balance tests to assess whether a treatment variable is assigned "as-if" randomly. However, balance on measured covariates is neither necessary nor sufficient to ensure a critical condition for causal inference: independence of treatment assignment and potential outcomes. We use an important debate over the randomness of close elections to clarify the limitations of existing balance tests and focus on a key determinant of their power and specificity—the extent to which covariates are predictive of potential outcomes. We prove that when covariates are sufficient for potential outcomes in a sense we define, covariate imbalance implies that as-if random fails. We then propose an omnibus "informativeness-weighted" balance test which weights covariate differences of means by measures of their prognosis. This approach limits false negatives when potential outcomes are imbalanced while minimizing spurious rejections due to imbalance on irrelevant covariates, as we show through simulation. We introduce an open-source software package that implements the test.

Keywords: Balance tests, natural experiments, regression discontinuity, prognostic covariates, prognostic scores, ignorability, weights

Acknowledgments: We are grateful to Lily Medina for helpful assistance and to Avi Feller, Ben Hansen, Erin Hartman, Christopher Harshaw, Sam Pimentel, Fredrik Sävje, Tara Slough, and participants at the 2022 LatAm PolMeth conference in Buenos Aires and the UC Berkeley Methods Workshop for valuable comments and suggestions. Prepared for presentation at the PolMeth conference, Stanford University, July 10-11, 2023.

[*]University of California, Berkeley
[†]University of California, Berkeley
[‡]University of California, Berkeley

1 Introduction

Methodologists often urge researchers to test observable implications of assumptions needed for causal inference. In natural experiments, regression-discontinuity designs, and related methods, a common approach is to test for statistical balance on pre-treatment covariates across treatment and control groups. The logic appears straightforward: since covariate values are determined prior to treatment assignment, then if a coin flip determined treatment assignment—as assumed or as stipulated by design in such studies—we would expect equal distributions of all covariates in the two groups.[1] Thus, up to chance error, we should find proportionately as many men as women, or as many young as old people, assigned to treatment and control conditions. A statistically insignificant association between treatment assignment and gender, age, or other covariates is consistent with "as-if" random assignment of a treatment, while the presence of such an association may suggest a flaw in the design.[2]

Unfortunately, such tests may shed no light on a key condition for causal inference: the independence of treatment assignment and potential outcomes, that is, the outcomes that would be realized under counterfactual assignment to different treatments.[3] Self-selection processes that lead to failures of as-if random may imply imbalances on some covariates but not on others. In an observational study of the efficacy of a new drug, sicker patients may select into the treatment group—but men may be as likely to do so as women. This leads to expected balance on gender across the treatment and control groups but imbalance on prior health. These covariates differ, however, in their informativeness about potential outcomes: health status after an intervention is likely tightly related to prior health status while gender may be unrelated to potential outcomes. If we only have data on gender, we may fail to reject random assignment; and yet potential health outcomes under assignment to treatment or control are very likely imbalanced. Conversely, men might tend to select into treatment, yet gender is unrelated to potential responsiveness to a treatment. This may lead us to reject as-if random based on the gender imbalance, even if potential outcomes are balanced across the treatment groups. Even if we have data on both gender and prior health, standard procedures such as covariate-by-covariate or (unweighted) multivariate balance tests do not capture well the asymmetry in the informativeness of these covariates. Thus, standard balance tests tests may reject spuriously, due to imbalances on irrelevant covariates unrelated to potential outcomes. Or they may fail to reject when potential outcomes are in fact imbalanced, because they do not take sufficient account of the extent to which covariates are prognostic—that is, predictive of potential outcomes.

We show in this paper how to diagnose and increase the power of balance tests to detect the dependence of treatment assignment on potential outcomes. When covariates are sufficient for potential outcomes, in a sense we define, finding imbalance on covariates implies imbalance of potential outcomes. However, tests using irrelevant covariates—those unrelated to potential outcomes—can lead researchers to over-reject as-if random when it is true or under-reject it when it is false. Projecting potential outcomes onto covariates before conducting balance tests helps to avoid this problem. A key insight is that post-treatment data on outcomes are often available at the time researchers conduct balance tests. This implies that the extent to which covariates are predictive of potential outcomes can be assessed empirically, for example, by using control- or treatment-group samples to estimate a finite population relation between covariates and potential outcomes.

[1] Covariates could also be defined to include "placebo outcomes," i.e., post-treatment variables known or assumed to be unaffected by treatment assignment (Eggers et al. 2021, Caughey et al. 2017).

[2] See e.g. Freedman (1999); Rosenbaum (2002, 2010); Hansen and Bowers (2008); Sekhon (2009); Imai et al. (2010); Dunning (2012); Caughey et al. (2017); or Eggers et al. (2021).

[3] Neyman et al. (1923), Rubin (1974), Holland (1986).

We develop a prognosis-weighted (alternately, "informativeness-weighted") test to assess the independence of treatment and potential outcomes.[4] The principal test statistic is based on the difference between the fitted value of the average $\widehat{Y_i(0)}$, the covariate-adjusted average potential outcome under control, in the treatment and control group samples. The statistic is equivalent to a weighted sum of the differences of means for each individual covariate, where the weights are the fitted coefficients from the standardized regression of potential outcomes on covariates in the control group sample. This omnibus test—"omnibus" because unlike covariate-by-covariate tests, it is based on the joint distribution of covariate differences—thus takes account of the "importance" of each covariate, or the extent to which different covariates are linearly prognostic. The test downweights irrelevant, non-prognostic covariates and upweights variables we would expect to be imbalanced if treatment assignment in fact depends on potential outcomes.

Our theoretical and simulation results suggest several key advantages of this approach. First and most importantly, the test is both more specific and more powerful than existing approaches. That is, it tends to fail to reject as-if random assignment due to imbalances on irrelevant covariates more often than existing approaches (and is thus more specific); but it will boost rejection probabilities relative to standard tests when prognostic covariates are imbalanced (and thus it is more powerful). Methodologists have rightly pointed out that balance tests are often poorly powered to reject false null hypotheses stipulating as-if random assignment (Cattaneo et al. 2015). Moreover, failing to reject a null hypothesis of as-if random assignment is not the same as accepting it, leading some methodologists to recommend alternatives to standard balance tests (Hartman and Hidalgo 2018; Hartman 2021).[5] Our results suggest a subtle relationship between the prognostic value of covariates and the power of tests, however. Unweighted omnibus tests may overreject null hypotheses when treatment assignment is independent of potential outcomes, because they give too much weight to irrelevant covariates unrelated to potential outcomes. Our weighted procedure can reduce false positives or Type I error in this case. However, when treatment assignment does depend on potential outcomes and we have prognostic covariates availables, our prognosis-weighted approach rejects as-if random more often than unweighted approaches, thus limiting false negatives or Type II error and increasing power. In other words, the test limits both false negatives and false positives.

Second, unlike covariate-by-covariate tests, our approach provides a clear rejection rule based on a single test statistic. Researchers often appear to rely on an informal rule of thumb in assessing the results of balance tests. For example, if a treatment is randomly assigned, we would expect significant covariate imbalances at the 0.05 level in only 1 out of 20 or 5% of independent tests. However, when making multiple statistical comparisons across different covariates and when tests are dependent—which occurs whenever covariates are correlated with each other, that is to say, almost always in practice—such a rule-of-thumb is not reliable. It can therefore be a matter of opinion whether the totality of the evidence from a set of covariate balance tests undercuts a claim of as-if random. We thus add to recent work that proposes the use of omnibus statistics or combinations of p-values, including those that allow for dependence among covariates (Hansen and Bowers 2008; Caughey et al. 2017; Gagnon-Bartsch and Shem-Tov 2019).

Finally, our approach provides a basis for assessing the evidentiary value of balance tests. Pre-treatment (lagged) measures of outcome variables tend to be highly prognostic, leading methodologists to counsel their use in falsification tests (Imbens and Rubin 2015: 483-4). Yet such predictive covariates may or may not be available. It is also an empirical question whether a lagged outcome or any other covariate is in fact prognostic in any study, as our motivating example in the next section suggests. In some settings *all* available covariates may be only weakly related to potential outcomes. Our approach suggests

[4] We offer an accompanying R package `pwtest` that implements the test we propose (forthcoming).

[5] See also Imai et al. 2008, who develop important critiques of balance tests that differ from the issues we raise in this article.

diagnostics that can help researchers assess the strength of the test. When none of the pre-treatment covariates to which researchers have access are prognostic, balance tests may have especially weak power over an alternative hypothesis that potential outcomes are imbalanced. Their results should then not be taken as strong evidence in favor of the key identifying assumption for causal inference.

We take inspiration from a large literature on multiple testing in statistics and epidemiology, which recommends upweighting tests for "important" hypotheses—or those that are most plausibly false—in p-value combinations.[6] However, our approach gives specific content to which hypotheses are most likely to be false in balance tests, by upweighting covariates that are related to potential outcomes. We also recommend constructing an omnibus p-value by evaluating the weighted sum of differences of covariate means directly, rather than by combining p-values from separate covariate-by-covariate tests. Caughey et al. (2017) develop a valuable non-parametric combination approach to generating omnibus p-values for placebo tests, including balance tests, using a combination metric owing to Fisher (1935). However, the choice of combination metric can introduce discretion; and this choice is not required for balance tests, in which mean differences for different covariates can be readily combined. Our approach is related to papers by Hansen and Bowers (2008), who develop a procedure for conducting balance tests based on combining covariate differences of means in block- (and cluster-) randomized experiments; and Gagnon-Bartsch and Shem-Tov (2019), who provide a classification permutation test. None of these important papers, however, considers variation in the prognostic power of different covariates. While many works on causal inference mention the usefulness of conducting tests for balance on prognostic covariates, the rationale is not always explicit, nor are the potential gains of doing so in terms of increased statistical power against a clear null hypothesis. Hansen (2008) proposes a "prognostic score" that is akin to our measure of prognosis, but he develops this approach for purposes of covariate adjustment in an outcome model. Our work is perhaps closest to Stuart et al. (2013), but here we develop and evaluate the properties of a novel test statistic to assess prognostic balance.[7] Our innovation is our focus on the association between covariates and potential outcomes as the basis for balance testing and especially our proposed test procedure.

In the next section, we discuss motivation for our focus on covariate prognosis, using as a case study a recent debate over the randomness of close elections as well as a sample of published natural experiments. Section 3 discusses why prognosis matters for balance tests, defining as-if random and presenting a theorem regarding a condition under which imbalance on covariates implies imbalance of potential outcomes. Section 4 develops statistical theory behind our informativeness-weighted test. In Section 5, we present simulation evidence on its power and specificity. We conclude by discussing several possible extensions.

2 Case study: are close elections random?

In an important study, Caughey and Sekhon (2011) appraise the use of regression-discontinuity designs to study the effect of incumbency, taking data from close U.S. House elections (1942-2008). A priori, in a very close election, which party winds up with a slightly greater vote share at time t seems quite plausibly as-if random. If so, assignment of the treatment—party incumbency—is independent of potential outcomes, as well as pre-treatment covariates, facilitating study of the impact on electoral outcomes at time $t + 1$. For this reason, the close-election design has become extremely widespread.

[6] Examples include Holm (1979); Benjamini and Hochberg (1997); and Genovese et al. (2006). See also Fisher (1935); Kost and McDermott (2002); or Westfall (2005).

[7] In Stuart et al. (2013) as well as Leacy and Stuart (2014) and Wainstein (2022), the focus is on adjustment for prognostic variables in estimation of an average treatment effect.

Caughey and Sekhon, however, use a series of covariate balance tests to assess local randomization in the neighborhood of the running variable determining treatment assignment, i.e., in a small bandwidth or window around the 0% vote margin.[8] Concerningly, several p-values from balance tests (their Figure 2) suggest statistically significant imbalances in past incumbency, as well as the winning party's past vote share, campaign spending, and measures of candidate quality, suggesting a possible failure of as-if random in very close elections.[9] Still, districts barely won by Democrats at time t do not differ from those barely won by Republicans on several other political and demographic variables, e.g. whether the state has a Democratic governor or secretary of state, the margin of victory in the presidential race, voter turnout, whether the seat is open, and the percentage of urban, Black, or foreign-born residents. Thus, we see imbalances on some covariates but lack of imbalance on several others.

How should one interpret the totality of the evidence in a balance plot such as Caughey and Sekhon's? These researchers (rightly, in our view) attribute particular importance to the imbalance on the winning party's past incumbency and vote share at time $t-1$, since these variables are presumably highly correlated with future incumbency and vote share. However, there is no formal procedure that takes into account the extent to which covariates are prognostic; and the covariates included in the balance tests are correlated, so simply comparing the number of rejections to the number of tests (e.g. to see if the ratio is greater than 1 out of 20) is not informative.

In a subsequent study, Eggers et al. (2015) confirm that lagged incumbency seems to be the major driver of imbalances in Caughey and Sekhon's data: in a procedure conceptually related to one we propose in Section 4, they show that Democratic near-winners and near-losers are not significantly different on pre-treatment covariates other than lagged incumbency, once the latter is controlled in a regression. They then extend the Caughey and Sekhon study to a broad range of majoritarian elections around the world, comparing close election winners and losers *only* on a measure of lagged incumbency. They find balance on this covariate in every other setting they examine, leading them to argue that the observed imbalance in U.S. House elections in the latter part of the twentieth century is unusual and may reflect special features of that context or may simply be due to chance.

The extent to which lagged incumbency is prognostic thus appears critical to adjudicating this debate about whether close elections are as-if random. The same is true of pre-treatment covariates in many other natural experiments or discontinuity designs. However, these studies and many others do not take prognosis into account formally or empirically. Eggers et al. (2015) are right to assess covariate balance across a wide range of elections. Yet they effectively assert that lagged incumbency is the only important covariate on which to test for balance across these contexts.[10]

In fact, the prognostic value of lagged incumbency varies across countries and types of elections. As we show in Tables 1-2 in the Online Appendix Section 7.1, in the Eggers et al. data, the correlation between the vote share of the incumbent party at time $t-1$ and time t is 0.79 across all countries and election types but varies from a low of 0.09 in Brazilian mayoral elections to a high of 0.91 in the German Bundestag (full data set); in close elections (defined by a bandwidth of 0.5, i.e., the margin between the

[8] Subsequent work on regression-discontinuity designs has debated whether analysts should rely on local randomization or a weaker assumption of continuity of potential outcome regression functions; we do not enter that debate here, but we follow the practice in this study of using covariate balance tests based on differences of covariate means.

[9] Caughey and Sekhon (2011) also present evidence of "sorting" at the 0% vote margin: the incumbent party wins the closest elections more than twice as often as it loses (see their Figure 1).

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

winning and runner-up party is less than 1 percentage point), it varies from a high of 0.32 in New Zealand's post-war parliament to a low of −0.16 is the Canadian House of Commons (1867-1911).[11] Perhaps most importantly, the average correlation is just 0.02 across all close elections studied by Eggers et al., while it is substantially higher in the post-war U.S. House elections studied by Caughey and Sekhon (0.83 in the full data and 0.24 in close elections).[12] We return in the next section to discussing why this variation in covariate prognosis matters for testing as-if random.

The prognosis of covariates is rarely considered systematically in balance testing. We coded a random sample of 150 articles using randomized experiments, natural experiments, and regression-discontinuity designs in three top journals in political science (the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*), stratifying by journal, over the time period 2000-2019.[13] Overall, 52 percent of the sampled articles presented covariate balance tests in the body or Appendix of the paper. The majority (56 percent) of those use only covariate-by-covariate tests, rather than some omnibus test statistic (such as the p-value for the F-statistic from the regression of a treatment indicator on all covariates). Only 18 percent of tests used a lagged dependent variable as a covariate. And we found no examples of systematic efforts to account for the prognostic importance of covariates in balance tests, for instance, using a weighted procedure like that we propose in this paper. The situation does not appear dissimilar in economics or other social science disciplines; in particular, procedures that take account of covariates' degree of prognosis appear absent in the applied literature.

The covariates used in balance tests vary substantially, nonetheless, in the extent to which they predict potential outcomes. We took a small further random sample from the 150 studies we coded, excluding the randomized experiments, stratifying by natural experiment versus discontinuity and on the presence of a lagged dependent variable or explicit discussion of prognosis in the paper. We had to exclude some natural experiments either due to lack of appropriate replication data or other considerations outlined in Online Appendix Section 7.2. We then calculated two measures for each included study: the multiple R^2 from the regression of observed potential outcomes in the control group on all available covariates ("Prognosis") and the multiple R^2 from the regression of a treatment assignment indicator on all available covariates ("Imbalance"). We note that we do not intend these measures as providing formal tests of prognosis or imbalance; we elaborate approaches to testing in the next section.

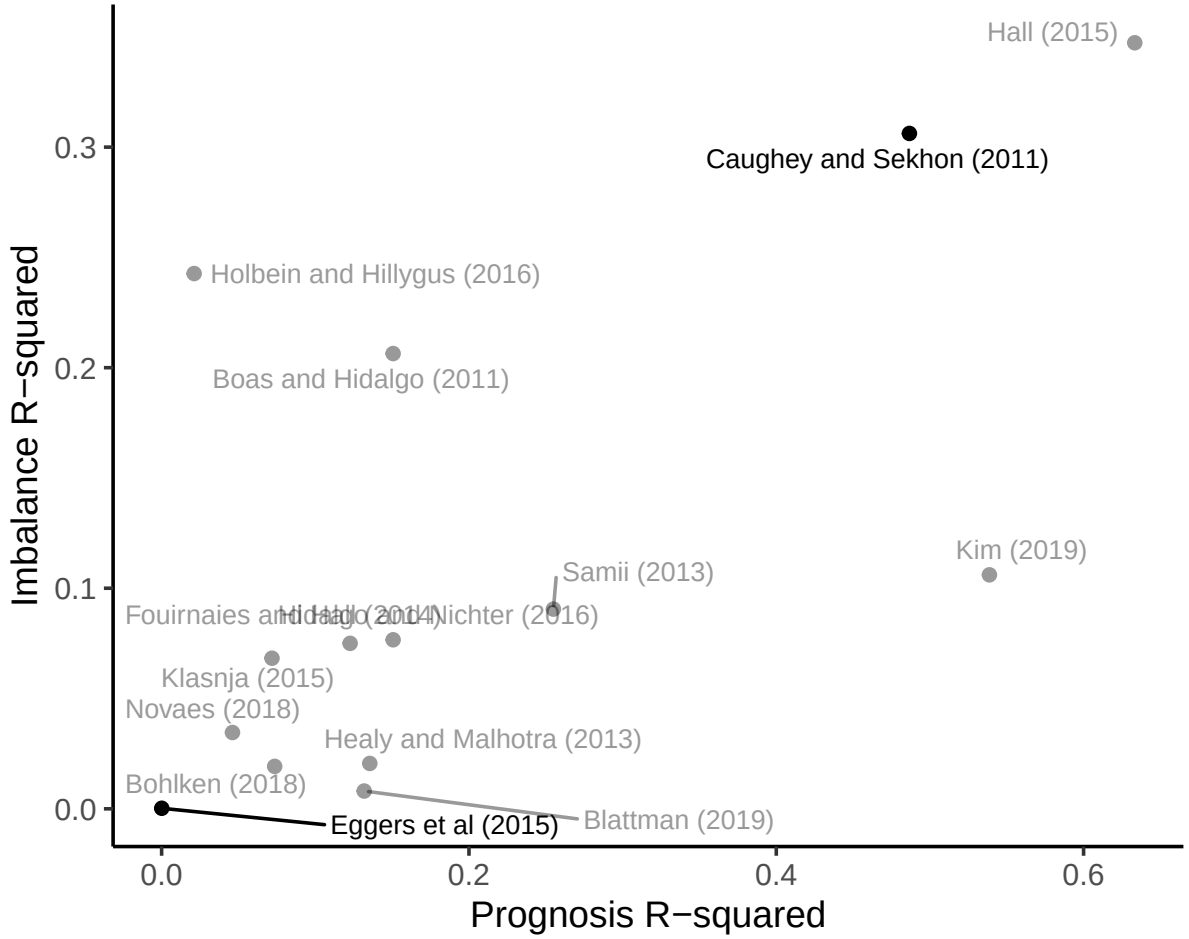
Figure 1, which plots these measures for our smaller sample of studies, suggests several insights. First, there is considerable variation across studies in the extent to which the covariates used in balance tests are prognostic. Close to the vertical axis, covariates are non-prognostic and thus bear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

[11] This comports also with findings on the varied causal effects of incumbency across contexts; see e.g. Schiumerini (2015).

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

[13] See Online Appendix Section 7.2. For code used in the sampling, see https://github.com/lilymedina/JSTOR_query.

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



The figure plots a random sample of natural experiments and regression-discontinuity (RD) designs drawn from all those published in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics*, 2000-2019; Caughey and Sekhon (2011) is added. Prognosis is the R^2 from a regression of potential outcomes under control on all available covariates (control group only). Imbalance is the R^2 from a regression of treatment assignment on all available covariates. See Appendix Section 7.2 for information on sampled studies and construction of the measures.

For reasons we develop in the next section, this variation in prognosis—as well as imbalance—is critical for forming persuasive tests of as-if random. Heuristically, there are four kinds of cases to consider. (1) Studies located in the upper-left quadrant of Figure 1 may be prone to spurious rejection with standard procedures—because there is imbalance on covariates unrelated to potential outcomes. (2) In the lower-left quadrant, the important concern is instead that none of the measured covariates are prognostic of potential outcomes but we find balance on treatment assignment—leading to a form of Type I error in which we fail to reject as-if random, yet potential outcomes themselves may be imbalanced. The Eggers et al. (2015) study is located here. (3) In the lower-right quadrant, we find cases with high prognosis but low imbalance: here, the claim of as-if random may be most persuasive. (4) Finally, in the upper-

right quadrant, rejection may be most persuasive of a failure of as-if random—because covariates are as a whole prognostic of potential outcomes. The Caughey and Sekhon (2011) study is located here. We note, however, that location in this quadrant need not imply rejection using the prognosis-weighted procedure we propose in the next section: covariates may be predictive of potential outcomes as a whole, leading to a high prognosis R^2, and yet imbalance may occur on a non-prognostic subset of covariates. We return to this logic later, after developing the reasons why covariate prognosis matters for assessing as-if random in the next section.

3 Are potential outcomes balanced? Why prognosis matters

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

First, the most direct test of the key identifying condition for causal inference would assess balance in *potential outcomes* across the treatment and control groups. A direct test of this assumption is impossible, however, due to the fact that once treatment has occurred, we do not observe potential outcomes under control in the treatment group or potential outcomes under treatment in the control group. A covariate that is strongly associated with potential outcomes may nonetheless give us substantial information about this realized balance. This may be the reason that methodologists advise sometimes balance tests on lagged dependent variables (e.g. Imbens and Rubin 2015: section 21.3), but the rationale for doing so is not always explicit—and as our example in Section 2 showed, lagged outcomes are not always prognostic.

Second and relatedly, theories of self-selection bias suggest that if as-if random assignment fails, covariates strongly related to potential outcomes are the most likely to be imbalanced across treatment and control groups. Indeed, if subjects have the opportunity to select into treatment and control groups, as in many observational studies, then—contrary to the assumption of as-if random—they may do so in a way that reflects the outcomes they would experience under treatment or control. For example, sicker patients may select into the group that receives a new treatment if given the chance, since they need the treatment. This leads us to expect imbalance on prior health if as-if random fails but not necessarily other covariates. As we will outline in Section 4, prioritizing such prognostic covariates when assessing balance allows more sensitive and more powerful tests of as-if random.

In the rest of this section, we show formally why the prognosis of covariates matters for testing as-if random. We first define the key identifying condition: treatment assignment is independent of potential outcomes. We then show that contrary to what is implied by standard covariate balance tests, statistical independence of covariates and treatment assignment does not imply that this condition holds. Nor does statistical dependence of covariates and treatment assignment imply that it fails. However, when covariates are sufficient for potential outcomes in a specific sense, covariate imbalance does implies imbalance in potential outcomes. Intuititively, covariates that are highly prognostic—and thus substantially informative about potential outcomes—may allow for a direct test of the key identifying condition.

3.1 Defining as-if random

We develop our argument using a design-based, finite population set-up. Consider a study with a completely enumerated finite population of N units indexed by $i = 1, \ldots, N$. 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.[14] 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.[15] 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 $n_1 + n_0 = N$; without loss of generality, but for ease of exposition, suppose $n_1 = n_0$.

A key condition in natural experiments, sufficient for identification of the ATE, is

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

where $\perp\!\!\!\perp$ denotes "is independent of."

In words, treatment is assigned independently of potential outcomes.[16] This ensures, for example, that sicker patients do not go systematically to the treatment group in a drug trial studying health outcomes, or that those less prone to vote do not disproportionately receive a vote-mobilizing intervention. If as-if random holds, the true treatment effect is estimable using simple, transparent methods: for example, a simple difference of mean outcomes in treatment and control groups is unbiased for the ATE (Freedman 1999). Assumption 1 is also sometimes called (strong) ignorability.[17]

Assumption 1 cannot be directly verified due to the the "fundamental problem of causal inference": $\{Y_i(1), Y_i(0)\}$ is not completely observed for any unit (Holland 1986). In a randomized experiment, independence of potential outcomes and treatment assignment is an implication of a chance protocol, often under the control of a researcher (Fisher 1933, Zhao and Ding 2021). In natural experiments, as-if random is held to be an implication of a concrete process that produces a haphazard allocation to treatments, in particular, one that does not depend on units' potential outcomes (Freedman 1999, Dunning 2012). With additional assumptions discussed below, however, it is possible to make Assumption 1 falsifiable.

3.2 Standard balance tests, and two counterexamples

3.2.1 Preliminaries

We consider a set of possible covariates $\mathscr{X}$. We suppose that $\mathscr{X} = \{\mathscr{X}^S, \mathscr{X}^N\}$ ('signal' and 'noise', respectively – compare Liu and Ruan 2020), where $\mathscr{X}^S$ contains information about potential outcomes and $\mathscr{X}^N$ does not. Treat $\mathscr{X}^S$ and $\mathscr{X}^N$ as finite but potentially unobserved. With a slight abuse of notation,[18]

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

[14] Neyman et al. (1923) or Rubin (1974).

[15] This formulation of potential outcomes embeds an assumption of non-interference or the stable unit treatment value assumption (Cox 1958, Rubin 1978), as in many formalizations of average treatment effects.

[16] Thus, there are $\binom{N}{n_1}$ possible vectors Z in which n_1 units are assigned to treatment and the remaining n_0 units go to control; each vector is equally likely and the chances do not therefore depend on the vectors $\{Y(1), Y(0)\}$.

[17] This can be contrasted with "weak ignorability," or $Z \perp\!\!\!\perp \{Y(1), Y(0)\}|X$, which we consider later.

[18] We say slight abuse of notation because $\mathscr{X}^S$ and $\mathscr{X}^N$ are fixed, as are potential outcomes $\{Y(1), Y(0)\}$; the notation $\perp\!\!\!\perp$ is often reserved for statistical independence of random variables. In equation (1), the symbol can be taken to mean that $\mathscr{X}^N$ is uncorrelated with $\{Y(1), Y(0)\}$ in the finite population.

A researcher measures a set of covariates $\boldsymbol{X} \subseteq \mathscr{X}$. (We use boldface for this matrix). These observed covariates may contain some, all, or none of the available information about potential outcomes.[19] If the data include the signal covariates, we say that $\mathscr{X}^S \subseteq \boldsymbol{X}$; if the data include the noise covariates, we say that $\mathscr{X}^N \subseteq \boldsymbol{X}$. If we measure *only* signal or noise covariates, we have $\mathscr{X}^S = \boldsymbol{X}$ or $\mathscr{X}^N = \boldsymbol{X}$, respectively.

There are two facts to notice about this setup. First, treatment assignment Z will be independent of $\mathscr{X}^S$ if and only if it is independent of potential outcomes.

Proposition 1 (Sufficiency of $\mathscr{X}^S$)**.** $Z \perp\!\!\!\perp \mathscr{X}^S \iff Z \perp\!\!\!\perp \{Y(1), Y(0)\}$

That is, because $\mathscr{X}^S$ contains all and only the information about potential outcomes, treatment assignment must be either independent of both $\mathscr{X}^S$ and potential outcomes, or not independent of both.

Second, the noise covariates $\mathscr{X}^N$ may or may not be associated with treatment assignment, but they contain no information about potential outcomes. This is nuisance information. Thus, we might observe covariates that are correlated with the treatment assignment process but of no relevance in predicting treatment effects. Or we might observe pure noise that is unrelated to both treatment assignment and potential outcomes.

With these preliminaries, we can now evaluate the standard logic of balance testing.

3.2.2 The Logic of Balance Testing

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

Claim 1. *(Standard Practice: Balance tests)*

Standard practice seems to imply that

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

Hence, $Z \not\perp\!\!\!\perp \boldsymbol{X} \iff Z \not\perp\!\!\!\perp \{Y(0), Y(1)\}$.

Claim 1 is not correct, however.

Example 1. *(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 $\Longleftarrow$ direction of Claim 1 does not follow.

This situation corresponds to one example in the introduction: perhaps men select into the treatment group in a drug trial, but gender is unrelated to potential responsiveness to treatment. We then expect statistical dependence between treatment status and gender—but this does not imply that potential outcomes are imbalanced.

[19] For the purposes of this section, we consider the case where covariates contain all or none of the information about potential outcomes. We study the case where covariates contain some of the information about potential outcomes by simulation, below.

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, if the imbalanced covariates are unrelated to potential outcomes, their imbalance does not constitute evidence that as-if random assignment as defined in Assumption 1 fails.

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

Example 2. *(Counterexample to Claim 1: False negatives)*

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

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

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

3.3 The informativeness of covariates

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

Definition 1. *([Minimal] Sufficiency of Covariates)*

A set of covariates $\boldsymbol{X} \subset \mathcal{X}$ is sufficient for $Y(1), Y(0)$ if

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

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

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

In words, if the observed covariates are sufficient for the potential outcomes, then they contain all possible observable information about potential outcomes. Moreover, if the covariates are minimal sufficient, then they contain all *and only* the possibly observable information about potential outcomes (and any smaller subset $\boldsymbol{S}$ of $\boldsymbol{X}$ would no longer be sufficient).[20]

When measured covariates are minimally sufficient, standard balance tests allow us to infer whether or not Z is assigned as-if at random. The intuition: if $\boldsymbol{X}$ is sufficient for the potential outcomes, then it must contain all the information contained in $\mathcal{X}^S$—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

[20]Equivalently, if $\boldsymbol{X}$ is sufficient, $\sigma(\mathcal{X}^S) \subseteq \sigma(\boldsymbol{X})$; moreover, if $\boldsymbol{X}$ is minimally sufficient, $\sigma(\mathcal{X}^S) = \sigma(\boldsymbol{X})$. See Lemma 1 in Online Appendix Section 7.3. This is also equivalent to Pearl (1988a)'s notion of a Markov Blanket.

and Z. 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.

Theorem 1. *(The Logic of Balance Tests when $\boldsymbol{X}$ is Minimally Sufficient)*

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 Online Appendix Section 7.3.*

Theorem 1 clarifies one condition that allows us to interpret balance tests as an assessment of as-if random. The $\Longrightarrow$ direction controls *false positives*: when covariates are minimally sufficient, we have that a failed balance test implies a failure of as-if random. The $\Longleftarrow$ direction controls *false negatives*: when covariates are sufficient (here we do not need minimal sufficiency), then when treatment is not assigned independently of potential outcomes, we should expect a well-powered balance test to fail.

An objective of our approach, in the informativeness-weighted test we present next (section 4), is to empirically construct a minimally sufficient set of covariates, via a projection of potential outcomes onto covariates. As we show, the extent to which this reduces false positives and false negatives depends on the prognosis of the covariates X at our disposal.

3.3.1 Diagnosing sufficiency

Requiring sufficiency may set too high a bar for practice: it is akin to the assumption that we have measured all relevant information about potential outcomes via covariates, i.e., that we observe $\mathscr{X}^S$. Yet, we may evaluate sufficiency empirically using goodness-of-fit measures. For example, the prognosis R^2 (see Figure 1) from the regression of potential outcomes on covariates indicates (linear) goodness of fit.[21] A high R^2 indicates little residual variation in potential outcomes once we condition on covariates and thus may make sufficiency more plausible; an R^2 of 1 from the finite-population regressions of Y(0) and Y(1) on covariates implies sufficiency. Moreover, balance tests can be powerful or specific even if observed covariates contain only some but not all of the information about potential outcomes (see section 4), and the prognosis R^2 helps diagnose the extent to which they do. To be sure, while a high prognosis R^2 may suggest sufficiency, the condition could hold even with a low R^2: there may be simply be much variation in potential outcomes that is not associated with any covariates $\mathscr{X}^S$. Thus, despite a low prognosis R^2, potential outcomes are unrelated to covariates in $\mathscr{X}^N$ once we condition on $\mathscr{X}^S \subseteq \boldsymbol{X}$.

In some settings, the lagged (pre-treatment) value of the outcome may be identically equal to the potential outcome under control—that is, $X_i = Y_{i,t-1} \equiv Y_i(0)$ for all i. Then, sufficiency for $Y(0)$ holds trivially.[22] This may be why methodologists sometimes counsel the use of lagged dependent variables in falsification tests (Imbens and Rubin 2015: 483-4), though the reasons are rarely directly articulated. The identity also requires an assertion of temporal stability, which would be violated if, for instance, there are heterogeneous time trends in the outcome. Moreover, as our examples in section (2) suggested, it is an empirical question whether any given covariate, including a lagged outcome, is in fact prognostic.

To summarize, we have shown in this section that the informativeness of covariates regarding potential outcomes affects our ability to test as-if random. When covariates are minimally sufficient in the sense of Definition 1, testing the independence of treatment assignment and covariates allows us to test the

[21] Alternatively, we might consider an F-test, or use a distance metric like K-L divergence.

[22] It might sometimes happen that all pre-intervention units are exposed to a treatment, and a randomized intervention removes some of them from treatment status; in that case, the logic would be parallel but reversed, and $Y_{i,t-1} = Y_i(1)$ in that case.

independence of treatment assignment and potential outcomes. When sufficiency does not hold, however, balance tests may be prone to both false positives or false negatives.

We also note that sufficiency of $\boldsymbol{X}$ for $\{Y(0), Y(1)\}$ implies $Z \perp\!\!\!\perp \{Y(0), Y(1)\}|\boldsymbol{X}$.[23] This, of course, is equivalent to "weak ignorability" of treatment assignment. One might ask, if the assumption of weak ignorability held, why not simply adjust for covariates when estimating the ATE? One answer is that this requires analysts to choose an adjustment strategy, which can introduce discretion and dependence on the adjustment model. Strong ignorability, when plausible, instead allows simpler and more transparent estimation of the ATE—a key virtue of strong designs. Our focus here is on testing as-if random, rather than on covariate-adjusted estimation of treatment effects.

Another answer, however, is that sufficiency of observed covariates may not hold (and in fact may be unlikely to hold in real applications)—and yet as-if random does. Even when covariates are not sufficient, the logic of our argument suggests that we may be able to form more specific and powerful tests of as-if random by prioritizing individual covariates that are more prognostic, because such covariates bring us closest to the potential outcomes. We explore this conjecture next, presenting and assessing the properties of a global test statistic that upweights prognostic covariates.

4 An informativeness-weighted covariate balance test

We now turn to our prognosis-weighted (or "informativeness-weighted") balance test—so called because each covariate is upweighted or downweighted according to its degree of prognosis, or the extent to which it is associated with potential outcomes. We first present the key test statistic and then develop its theoretical rationale. The statistic can be written as

$$\delta_{PW} \equiv \sum_{j=1}^{p} \widehat{\beta_j^C} \delta_j. \tag{2}$$

with δ for "difference" and "PW" for prognosis-weighted. In (2), each δ_j is the difference of means on covariate j across the treatment and control groups—i.e., each is a standard test statistic in a covariate-by-covariate balance test. Each δ_j is then multiplied by the weight $\widehat{\beta_j^C}$, which is the jth coefficient from the (standardized) multiple regression of outcomes on covariates, as fit in the control group. Thus, equation (2) is the weighted sum of individual covariate differences of means, where the weights measure prognosis for potential outcomes under control.

Our approach improves on standard practice in two ways: (i) in contrast to covariate-by-covariate balance tests, δ_{PW} is single statistic to which we may attach a single p-value summarizing the strength of the evidence against as-if random; and (ii) we weight each covariate difference of means by an estimate of its importance—i.e. the extent to which it is linearly prognostic or predictive of potential outcomes under control. As we discuss later, some standard tests—such as F-tests from the regression of treatment assignment on all covariates—avoid the problem that (i) solves, but as unweighted tests they are still subject to the limitations that (ii) addresses.[24] We also note that our approach gives researchers incentives to gather a wide range of covariates and include them in their tests, since this improves the prognosis R^2 and other goodness-of-fit measures that can serve as diagnostics. In the rest of this section, we present statistical

[23] This can equivalently be stated as $\{Y(0), Y(1)\} \perp\!\!\!\perp Z|X$; see Dawid (1979).

[24] See e.g. Hansen and Bowers (2008) for discussion of other problems with F-tests.

underpinnings of the PW test statistic, derive its large-sample distribution, and present a resampling-based hypothesis test.

4.1 A finite-population regression

Consider the conditional expectation function $E(Y_i(0)|X_i)$ that gives the average value of potential outcomes under control at each value of $\boldsymbol{X}$ in the study group. The expectation is taken over a unit sampled at random from this finite population, or from each population stratum defined by a particular value of the vector $X_i = x$. The function gives the average of $Y(0)$ at each value of $\boldsymbol{X}$.[25]

If we observed $Y_i(0)$ for all units in the finite population, we could use a linear or a more flexible non-linear regression to approximate these conditional expectation functions. For example, the linear regression-adjusted value of $Y(0)$ given $\boldsymbol{X}$ in the finite population is

$$Y_i(0)_{lr} = X_i\beta, \tag{3}$$

where "lr" denotes the linear regression and i denotes any unit. Here, X_i is a $1 \times p$ vector of covariates, and β is the $p \times 1$ vector of coefficients from the finite-population regression. We could also approximate the conditional expectation function $E(Y_i(0)|X_i)$ using nonlinear or smoothed regressions, binning, or a range of classification procedures associated with machine learning. We emphasize however that our interest is in improving the performance (i.e. the power and specificity) of balance tests. The estimation of the conditional expectation function is not itself of direct interest; it plays a role only insofar as it shapes the performance of our test, and as we discuss later, non-linear classification or binning procedures may have some disadvantages relative to the linear regression in (3). Also, although one can define an analogous finite-population regression using values of $Y(1)$ (as we discuss further in the conclusion), we focus here on $Y(0)$ for a simple reason: in many natural experiments, we may observe pre-treatment values of the outcome in a "no treatment" status. Values of this variable may tend to be prognostic of potential outcomes under control $Y(0)$ but may or may not be prognostic of $Y(1)$. Moreover, mixing $Y(1)$ and $Y(0)$ values can require additional assumptions that may not be tenable (Hansen 2008).

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

$$\beta_j = \frac{\text{Cov}(Y(0), \widetilde{X_j})}{\text{Var}(\widetilde{X_j})}, \tag{4}$$

where $\widetilde{X_j}$ is the residual from the finite-population regression of X_j on the other $p - 1$ covariates.[26] When potential outcomes and covariates are standardized, each β_j is a standardized multiple regression coefficient.[27] More prognostic covariates will have larger absolute values of standardized β_j. Conversely, the coefficient β_j vanishes when the partial correlation between $Y(0)$ and X_j is zero.

[25] Here, $\boldsymbol{X}$ is a rectangular $N \times p$ data matrix where each row is the vector X_i. The elements X_{ij} denote the value for unit i on covariate $j = 1, \ldots, p$.

[26] This is due to the so-called Frisch–Waugh–Lovell (FWL) theorem, also called "regression anatomy;" see inter alia Angrist and Pischke 2009: 3.1.2 or Freedman 2009: Ex. 3.17.

[27] The standardized values of $Y(0)$ and X_j are $Y_i(0) - \overline{Y(0)}/\sigma_{Y(0)}$ and $X_{i,j} - \overline{X_j}/\sigma_{X_j}$, respectively, where $\overline{Y(0)}$ and $\overline{X_j}$ are the respective averages and $\sigma_{Y(0)}$ and σ_{X_j} are the standard deviations in the finite population.

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

4.2 Testing as-if random

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

$$\begin{aligned} H_0 &: E[\overline{Y(0)^T} - \overline{Y(0)^C}] = 0 \\ H_A &: E[\overline{Y(0)^T} - \overline{Y(0)^C}] \neq 0. \end{aligned} \tag{5}$$

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

To form a test of the null, the problem is then to estimate the unobserved $E(\overline{Y(0)^T})$. Consider first a regression of the outcome variable on covariates in the control group, i.e., the sample version of equation (3). We have exactly

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

where $\overline{X^C}$ (a $1\times p$ vector) gives the average value of the p covariates in the control group and the $p\times 1$ vector $\widehat{\beta^C}$ gives the coefficients from the control group regression.[28] Descriptively, the control group regression evaluated at the average value of the covariates is $\overline{Y(0)^C}$: a linear regression goes through the point of averages.

Moreover, equation (6) can also be viewed as a regression-weighted estimator for the average potential outcome under control in the finite population. The logic in brief: while $Y(0)_{lr}$ is incompletely observed—because we do not see $Y_i(0)$ for units in the treatment group and thus cannot fit equation (3)—the treatment and control groups are exchangeable under as-if random assignment. Viewed from another perspective, the control group is a simple random sample from the finite population, and so we can appeal to well-known sampling theory to form a consistent estimator (e.g., Cochran 1977, Chapter 7).[29]

[28] Thus,

$$\widehat{\beta^C} = (\sum_{i=1}^{n_0} X_i X_i^{'})^{-1} \sum_{i=1}^{n_0} X_i Y_i(0),$$

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.

[29] We show in the proof of Theorem 3 below that $\widehat{\beta^C}$ is a consistent estimator for β under as-if random.

We cannot run a regression analogous to equation (6) in the treatment group, however, because in that group we see potential outcomes under treatment, rather than potential outcomes under control. However, by the same logic of exchangeability, the expectation of the coefficient we would obtain—if we could run that regression in the treatment group—is clearly the same as the expectation 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

$$\widehat{\overline{Y(0)^T}} = \overline{X^T}\ \widehat{\beta^C}, \tag{7}$$

where $\overline{X^T}$ is the vector of average values of covariates in the treatment group. Here, we add a $\widehat{\ }$ over $\overline{Y(0)^T}$ because we are using $\widehat{\beta^C}$ to estimate the potential outcomes under control in the treatment group.

With an estimator of $\overline{Y(0)^T}$ in hand, we can form a statistic to test H_0 in (5). Subtracting (6) from (7), we have

$$\begin{aligned} \widehat{\overline{Y(0)^T}} - \overline{Y(0)^C} &= (\overline{X^T} - \overline{X^C})\,\widehat{\beta^C} \\ &= \delta_{PW}, \end{aligned} \tag{8}$$

where δ_{PW} is as defined in equation (2). The right-hand side of (8) is the prognosis-weighted sum of covariate differences of means. In the next sub-sections, we derive the conditional distribution of δ_{PW} and show how the statistic can be used to test H_0 with a resampling-based hypothesis test. We recommend standardizing $Y(0)$ and all covariates before running the regressions in equations (6) and (7) and forming the weighted sum δ_{PW}; this is the default option in our accompanying `R` package. This ensures that the contribution of each term to the sum is not a function of the measurement scale.

The logic suggests why we can extend our baseline approach to non-linear or smoothed regressions such as loess. Indeed, given data on X, we could use any classification procedure, such as random forests and other machine learning techniques, to fit $\overline{Y(0)}$ in the control group and use that fit to estimate $\overline{Y(0)}$ in the treatment group. We implement these methods as options in our accompanying R package. Such techniques may come at the cost, however, of the ready interpretation of coefficients as measures of the relative informativeness of different covariates, as in (4), though the variable importance metric in random forests is a possibility. We believe one attractive feature of our test statistic δ_{PW} is simplicity, as well as its connection to existing practice: it is simply a sum of the standard covariate differences of means but with each difference weighted by a measure of its (linear) prognosis for potential outcomes under control.

Under as-if random assignment, the probability limit of the random variable δ_{PW} is zero (see Theorem 3 below). Moreover, when $\boldsymbol{X}$ is sufficient for $Y(0)$ in the sense of Definition 1, rejecting H_0 implies rejecting as-if random (i.e. rejecting Assumption 1), as we show in Theorem 4 below. The upshot is that we have a single test statistic, δ_{PW}, that we can use to test H_0, as we will describe in sub-section 4.4. Rejections of the null will be due to imbalances in covariate means (i.e. $\overline{X^T} - \overline{X^C}$ in equation 8). However, in contrast to unweighted differences of covariate means—which may, for reasons we have explored, shed little light on imbalances of potential outcomes—the statistic δ_{PW} upweights prognostic covariates and downweights non-prognostic ones and thus better approximates the unobserved difference $\overline{Y(0)^T} - \overline{Y(0)^C}$.

4.3 Properties of the test statistic δ_{PW}

We derive here the large-sample distribution of our key test statistic δ_{PW} and relate it to (i) the distribution of an unweighted sum of covariate means and (ii) the distribution of another common unweighted statistic used for testing the multivariate equality of means, Hotelling's T^2. For testing, we recommend the resampling-based hypothesis test proposed in the next sub-section, for which an analytic expression for the variance of δ_{PW} is not required. Yet, comparing the weighted and unweighted measures is useful, since we assess the performance of tests based on these statistics in our simulations in section 5.

Consider as a benchmark the unweighted sum,

$$\delta_{UW} = \sum_{j=1}^{p} \delta_j, \tag{9}$$

where each δ_j is as in (2) and "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}.[30] 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.[31]

Theorem 2. *(Distribution of the unweighted sum of covariate differences of means) When treatment assignment is randomized, $E(\delta_{UW}) = 0$, and 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, δ_{UW} is asymptotically normal. ***Proof****: See Online Appendix subsection 7.4.*

Note that the variance in (2) 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.[32] However, unlike in that case, here there are no unobservable sample covariances because X_i is invariant to treatment assignment.[33] 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. The Online Appendix, Section 7.4, has further discussion.

The distribution of δ_{UW} 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

[30] 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.

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

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

[33] 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.

possible efficiency gains; see the Online Appendix section 7.5 for details. Hotelling's T^2 can in turn readily be related to the F-distribution used in some multivariate (unweighted) covariate balance tests.

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

Theorem 3. *(Distribution of the prognosis-weighted sum of covariate differences of means, δ_{PW}) When treatment is randomly assigned,* $\text{plim}(\delta_{\text{PW}}) = 0$. *The large-sample variance of δ_{PW}, conditional on the weights, is proportional to* $\text{Var}(\delta_{UW})$ *as given in Theorem 2.* ***Proof:*** *See Online Appendix sub-section 7.6.*

Theorem 3 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_{PW} to the square root of an estimator of the conditional variance referred to in theorem 3 (see the Online Appendix for details). However, such a large-sample approximation may not be reliable in small studies.[34] 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 resampling-based hypothesis test we develop in the next sub-section.

Note that 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. For example, let $\widehat{Y(0)_{lr}} = X^{C\prime}\widehat{\beta^C}$ be the control group regression (i.e., the sample version of 3). Then we have

Theorem 4. *(Observable Implication of As-If Randomization)*

Suppose that 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: $Z \not\perp \widehat{Y(0)_{lr}} \implies Z \not\perp Y(0)$. ***Proof****: see Online Appendix sub-section 7.7.*

This provides an empirical corollary to Theorem 1: when $\boldsymbol{X}$ is sufficient, we can validly use the prognosis-weighted statistic to test as-if random.[35] Moreover, as we show through simulation in Section 5, even when $\boldsymbol{X}$ is not sufficient the use of δ_{PW} improves the power and specificity of tests of as-if random. When covariates are substantially prognostic even if not sufficient in the sense of Definition (1), using the prognostic-weighted test avoids both false positives and false negatives. Thus, it improves our ability to use covariate balance tests to assess the independence of treatment assignment and potential outcomes.

4.4 A resampling-based hypothesis test

Randomization tests are widely used in experiments and natural experiments; they are attractive as they allow comparison of the observed value of a test statistic to its exact randomization distribution.[36] We propose here a variant that is a re-sampling (a.k.a. bootstrap) technique appropriate to our setting. It uses draws from the observed data to approximate the sampling distribution of δ_{PW}.

To illustrate the test, suppose we have a study with one treatment and one control group (i.e., treatment assignment is binary). Then the resampling test works as follows:

[34] 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).

[35] Note that δ_{PW} 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 Theorem 3.

[36] See Fisher (1935); also inter alia Caughey et al. (2017).

1. Draw a sample with replacement from the observed control group and regress (potential) outcomes on covariates in this sample; call the fitted regression coefficient vector $\widehat{\beta^{C*}}$. Also, calculate the sample mean of the covariates; call this $\overline{X^{C*}}$.

2. Take another independent sample, also with replacement and *also from the observed control group* and calculate the sample mean of the covariates; call this $\overline{X^{T*}}$. Putting together steps (1) and (2) allows us to calculate a simulated $\delta_{PW}^{*b} = (\overline{X^{T*}} - \overline{X^{C*}})'\widehat{\beta^{C*}}$.

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

4. Finally, calculate a two-sided randomization-based p-value as

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(|\delta_{PW}^{*b}| \geq |\delta_{PW}^{\text{obs}}|), \tag{10}$$

   where $\mathbb{1}$ is an indicator function that takes on the value of 1 if its argument is true and 0 otherwise. Reject the null if, e.g., $p^* < 0.05$.[37]

Several points about this procedure are useful to make. First, it validly simulates the null distribution of δ_{PW} when as-if randomization holds, for three reasons: (1) The expectation of the covariate difference of means is clearly zero when treatment assignment is randomized, as it is here: we are comparing the expected values of averages of two independent samples drawn from the same finite bootstrap population (Freedman 2009: Section 8.1).[38] (2) Critically, the procedure also allows in a natural way for the statistical dependence between the random variable $\widehat{\beta^C}$—as fit in the control group—and $\overline{X^C}$, with treatment assignment as the only source of stochastic variation. (3) Finally, note that the approach uses only $Y(0)$ values, rather than mixing $Y(0)$ and $Y(1)$ values as a permutation-based approach would do, which could lead to bias when $\boldsymbol{X}$ predicts $Y(1)$ differently than it predicts $Y(0)$.

Second, the resampling procedure can be adapted to accommodate a wide range of designs and modes of treatment assignment, for instance, clustered or blocked randomization. We include these as options in our accompanying software package.

Finally, 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 used during the "design stage" as well as for estimating average treatment effects (Rubin, 2007; Abadie and West, 2018; Liao et al., 2023). Here, we develop prognostic weights for purposes of balance testing, not for estimating average effects on outcome variables as in Hansen (2008).

5 The power and specificity of the informativeness-weighted test

We now turn to simulations to evaluate the performance of our approach. Even when covariates are not sufficient in the sense of Definition 1, our prognosis-weighted test may be more informative that standard

[37] Thus, we compare the absolute value of δ_{PW}^{obs} (or its unweighted version δ_{UW}^{obs}) to its randomization distribution and reject the null hypothesis if such an observed value would arise in fewer than, say, 5% of randomizations under the null.

[38] Note that in fact the treatment and control group means are dependent, and the two samples are drawn without replacement. Yet since X_i is the same whether unit i is assigned to treatment or control, it is as if the two samples were drawn independently, using a well-known argument found in Neyman (1923); see also Freedman (2007: A32-A34); Samii and Aronow (2011, Theorem 2); Gerber and Green (2012: 57); or Dunning (2012: 193).

covariate-by-covariate or (unweighted) multivariate balance tests. The potential advantages of projecting potential outcomes onto covariates before conducting a test is that (a) we heighten sensitivity to imbalance on covariates related to potential outcomes, thus limiting Type II error (reducing false negatives) and making the test more powerful; but (b) we can avoid rejections due to imbalance on irrelevant covariates, thus limiting Type I error (reducing false positives) and making test more specific. In contrast, we expect weaker power and specificity from unweighted tests, since they do not take unequal prognosis of covariates into account.

The extent to which these advantages are realized may vary across different data sets and data-generating processes, however. This makes the performance of our approach well-suited for investigation by simulation, to which we turn in this section. Two questions are key: (1) how prognostic must covariates be to test as-if random adequately? (2) How does the performance of the informativeness-weighted test compare to other possible approaches? Here we assess answers to these questions under different degrees of prognosis and covariate imbalance. We also compare the prognosis-weighted test to that using δ_{UW} in equation (9) as well as Hotelling's T^2. We consider settings in which covariates are minimally sufficient or merely sufficient and those in which sufficiency does not hold.

5.1 Evidence from simulations

Our simulations proceed in the following steps:[39]

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 $p = 1, 2$. 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. Potential outcomes under control are formed as a linear function of covariates such that $Y(0) = \beta_1 X_1 + \beta_2 X_2$, where β_1 and β_2 determine the prognosis of the corresponding covariate. The average treatment effect here is zero, i.e. $Y_i(0) = Y_1(1)$ for all i, but this plays no role in the simulation. The data-generating process therefore allows 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 δ_{UW}, δ_{PW}, and Hotelling's T^2.[40]

Step 3: (Resampling test). We conduct the resampling-based hypothesis test described in subsection 4.4 (with $B = 500$), calculating p-values from equation (10) for δ_{PW}. We also calculate analogous randomization p-values for δ_{UW} and Hotelling's T^2. 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 $p^* < 0.05$.

[39] Simulations were run on UC Berkeley's High Performance Computing server. The process outlined in Steps 1-6 runs in parallel on 24 CPU and takes on average 40 hours.

[40] We use the Hotelling package in R.

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.

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

Step 6: (Minimal sufficiency, sufficiency, and not sufficiency).

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_{PW}, δ_{UW}, δ_{PW}, 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 $cor(X_2, Z) = 0.15$.

The structure of the simulations allows us to compare the performance of our informativeness-weighted estimator to unweighted estimators (i) when as-if random is true and (ii) when it is false, i.e. treatment assignment depends on potential outcomes. Note that when $\text{Corr}(X_p, Z) \neq 0$, we have expected *imbalance* on X_p; when $\text{Corr}(X_p, Y(0)) \neq 0$, X_p is *prognostic*, with the correlation reflecting the degree of prognosis. Note that given the data-generating process in Step 1 above, whenever X_1 and X_2 are both balanced in expectation, the null of as-if random is true. This null also holds when one or both of these variables is non-prognostic, even if imbalanced in expectation. However, when at least one of X_1 and X_2 is imbalanced and prognostic, treatment assignment is associated with potential outcomes so as-if random fails and H_0 is false.[41] In this case, there would also be bias in a treatment effect estimator $\widehat{ATE}$ such as the unadjusted difference of means. A measure of Type I error—the false negative rate—is given by the rejection rate when as-if random holds. Type II error, on the other hand, constitutes failure to reject when the null is false. Statistical power is given by the proportion of rejections in the latter case.

Figures 2-4 show the results for Cases 1-3, respectively.[42] In every plot, the prognosis (true standardized coefficient) of X_1 varies along the horizontal axis. The vertical axis measures rejection rates (the proportion of the 1000 tests in which the null is rejected, given particular combinations of prognosis and imbalance parameter values). In each case, the prognosis of covariate X_2 (its true standardized coefficient)

[41] We implement a technical correction to ensure this holds in the simulation; see our replication code available online (forthcoming).

[42] Tables A3-A5 in the Online Appendix present a selection of the raw results.

is fixed at 0.25; in the first two cases (minimal sufficiency and sufficiency), this covariate is balanced in expectation (independent of treatment assignment), while in the third case it is imbalanced (dependent on treatment assignment) with $\text{Corr}(X_2, Z) = .15$ in expectation. The label at the top of each facet gives the expected imbalance, measured as the correlation ("rho") between treatment assignment and the covariate X_1. Each point in each facet thus depicts—for a particular degree of expected prognosis and imbalance—the proportion of rejections across the 1000 datasets generated with this particular correlation structure. We plot results for three test statistics: our prognosis-weighted sum δ_{PW}, the unweighted sum δ_{UW}, and Hotelling's T^2 (solid and dashed curves). In each figure, pink shading indicates those regions of parameter values where the null hypothesis is true.

First consider minimal sufficiency (Case 1, Figure 2). In the plots in the top panel of the figure, potential outcomes are in truth a linear function of X_1 and X_2 and we use both to construct the tests. Here X_2 is independent of treatment assignment (always balanced in expectation). Thus, in the top left facet, where X_1 is also balanced in expectation, as-if random is everywhere true regardless of the prognosis of X_1. All the tests perform well by failing to reject when the two sufficient variables are balanced (left panel). In the right panel, however, X_1 is imbalanced at $\rho = 0.1$, which corresponds to a standardized difference of means of about 0.2. Two aspects of the results are noteworthy. First, when the null of as-if random is false and as X_1 prognosis increases, the test becomes more powerful, surpassing the power of unweighted tests. Second, because the unweighted test is completely insensitive to prognosis, it rejects at the same rate no matter whether X_1 is prognostic or not. The flatness of the curves for the unweighted tests indicate their lack of sensitivity to covariate prognosis. But when X_1 prognosis is zero, it is an irrelevant covariate: the null is true and yet the unweighted tests fail to control false positives at an appropriate rate, whereas our proposed test correctly minimizes false positives in such cases.

Next, consider the case where observed covariates are sufficient but not minimally so (Figure 3). Thus, here we add a third covariate X_3, which is non-prognostic (irrelevant for outcomes) but may be imbalanced. When all three covariates are balanced in expectation (top-left plot) and the null is everywhere true, the unweighted tests and the prognosis-weighted test all control Type I error. However, when X_1 is imbalanced (top-right plot), the unweighted tests reject the null even when X_1 is non-prognostic and so as-if random is true (around 25 percent of the time in the case of the Hotelling test). Conversely, when X_1 prognosis increases above zero and so the null is false (given that it is also imbalanced, i.e. not independent of potential outcomes), the power of the prognosis-weighted test increases beyond that of the unweighted test. Here we see the prognosis-weighted test balancing specificity and power, as we expect.

In the bottom plots of the sufficiency case (Figure 3), the irrelevant covariate X_3 is highly imbalanced. In the bottom-left plot the null of as-if random is again everywhere true: the relevant (prognostic) covariates X_1 and X_2 are both balanced. But the unweighted test is sensitive to the imbalance on the irrelevant and imbalanced covariate X_3 and thus commits Type I errors at an elevated rate. In the final plot in the panel (on the bottom-right), the irrelevant covariate remains imbalanced but so is the relevant (potentially prognostic) covariate X_1. Thus, the null is false because of the imbalance on X_1. The unweighted test rejects at high rates (even higher than the prognosis-weighted test with these parameter values) but it does so because it is sensitive to imbalance on the irrelevant covariate. Indeed, the Hotelling test rejects over 80 percent of the time even when X_1 is not prognostic and so its imbalance leads to no imbalance of potential outcomes (i.e., the null is true). In contrast, the prognosis-weighted test is specific (fails to reject when the null is true) but its power (the rejection rate when the null is false) increases with prognosis.

As we discuss later, prognosis-weighted tests perform better in settings where observed covariates are good predictors of potential outcomes under control. The higher the R^2 of a regression of control potential

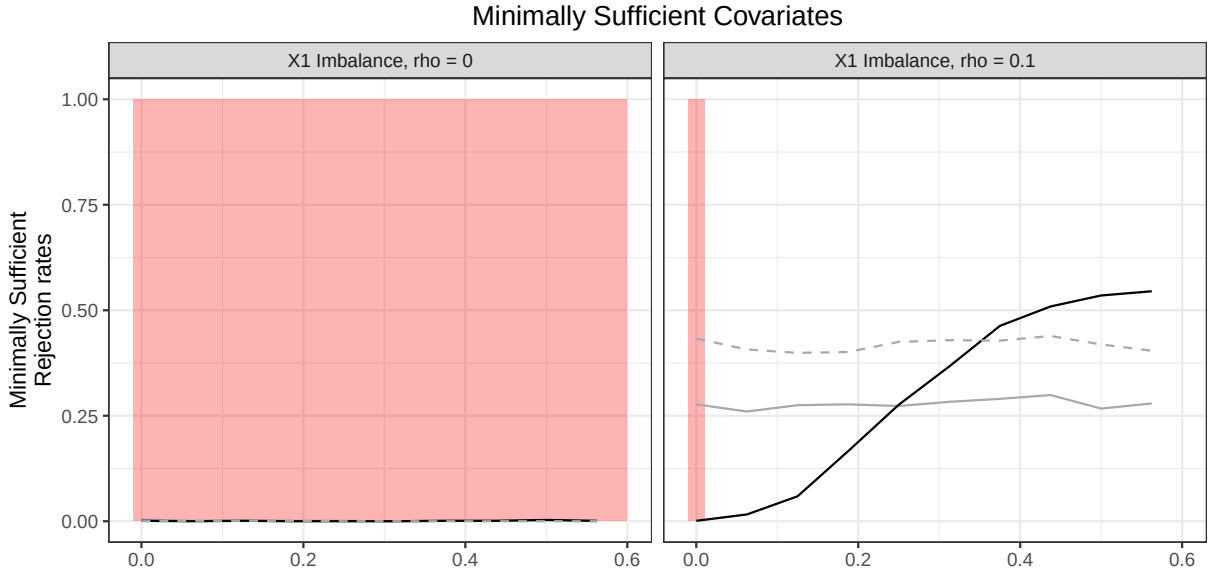


Figure 2: We simulate an $N = 500$ data set specifying expected correlations between covariates X_1 and X_2 and a treatment vector Z (covariate imbalance) and between each covariate and $Y(0)$ (covariate prognosis). We calculate the observed unweighted and prognosis-weighted sums of differences of covariate means and Hotelling's T^2 statistic and obtain randomization inference p-values for each statistic by permuting treatment assignment $B = 500$ times. We repeat this process 1000 times to obtain rejection rates for each set of expected correlations. The figure displays rejection rates as X_1 prognosis increases and for a case when X_1 is balanced (and thus the null hypothesis is everywhere true) and another in which it is imbalanced. X_2 is set to have a fixed prognosis of 0.25 and is always balanced in expectation. The shaded regions correspond to cases where as-if random holds: $Z \perp\!\!\!\perp \{Y(0), Y(1)\}$.

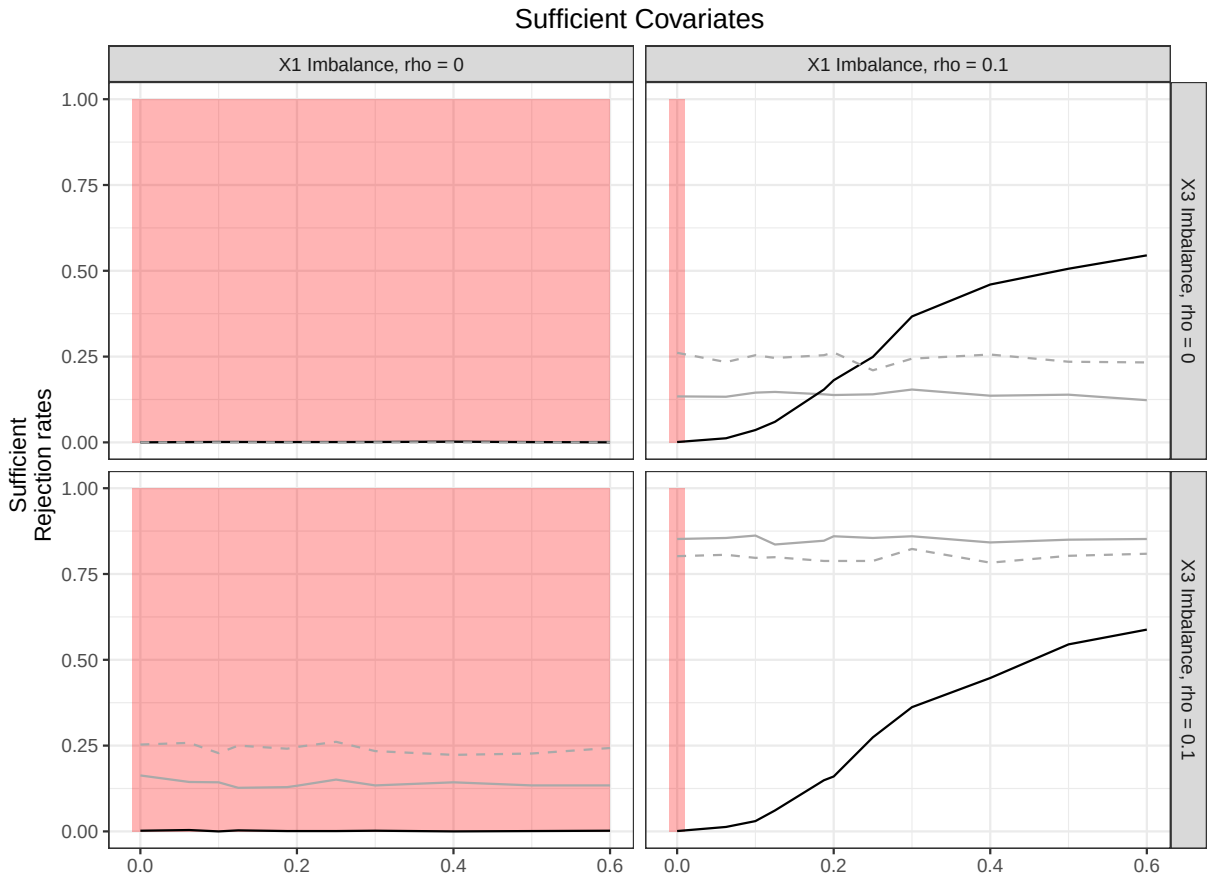


Figure 3: The simulation set up is the same as in Figure 2. However, in this case we consider tests that use signal covariates X_1 and X_2 as well as a noise covariate X_3 independent of potential outcomes. The figure displays rejection rates as X_1 prognosis increases and for cases where X_1 and X_3 are balanced and imbalanced. X_2 is set to have a fixed prognosis of 0.25 and is always balanced in expectation. The shaded regions correspond to cases where as-if random holds: $Z \perp\!\!\!\perp \{Y(0), Y(1)\}$.

outcomes on observed covariates, the less likely it is that we are in a setting similar to the one in the bottom-right plot.

Finally, consider the case in which observed covariates are not sufficient (Figure 4). Here, we only have access to the potentially prognostic covariate X_1 and the irrelevant covariate X_3; the prognostic and here imbalanced ($\rho = 0.15$) covariate X_2 is "omitted." The prognosis and imbalance of X_2 imply that the null is everywhere false in this figure: treatment assignment and potential outcomes are not independent. In the top-left plot, X_1 and X_3 are both balanced and so all tests reject at low rates, producing a false negative rate of nearly 1. This reflects the failure to measure the relevant and imbalanced covariate X_2. However, when X_1 is imbalanced, the tests pick up this imbalance. As X_1 becomes more prognostic (both in absolute terms and relative to the prognosis of X_2, which is fixed), the prognosis-weighted test becomes more powerful than the unweighted test, given balance on the irrelevant covariate (X_3). In the bottom two plots, where X_3 is imbalanced, we see (as in Case 2) that the unweighted test rejects at high rates (even higher than the prognosis-weighted test), but again it does so because it is wrongly sensitive to imbalance on the irrelevant covariate. Whereas the prognosis-weighted test rightly becomes more sensitive to imbalance on X_1 as prognosis increases, the unweighted test is invariant to covariate prognosis. The unweighted tests are insensitive to the null being true or false—they are sensitive only to covariate imbalance—whereas the rejection rate of the prognosis-weighted test is increasing in the evidence against as-if random.

While Figures 2-4 report results from one set of simulations, we reach similar conclusions from a broader set of simulations reported in Figure 5. Here, 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. 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 Figures 2-4, in Figure 5 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 2-4, 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 experimental studies in Figure 1, 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

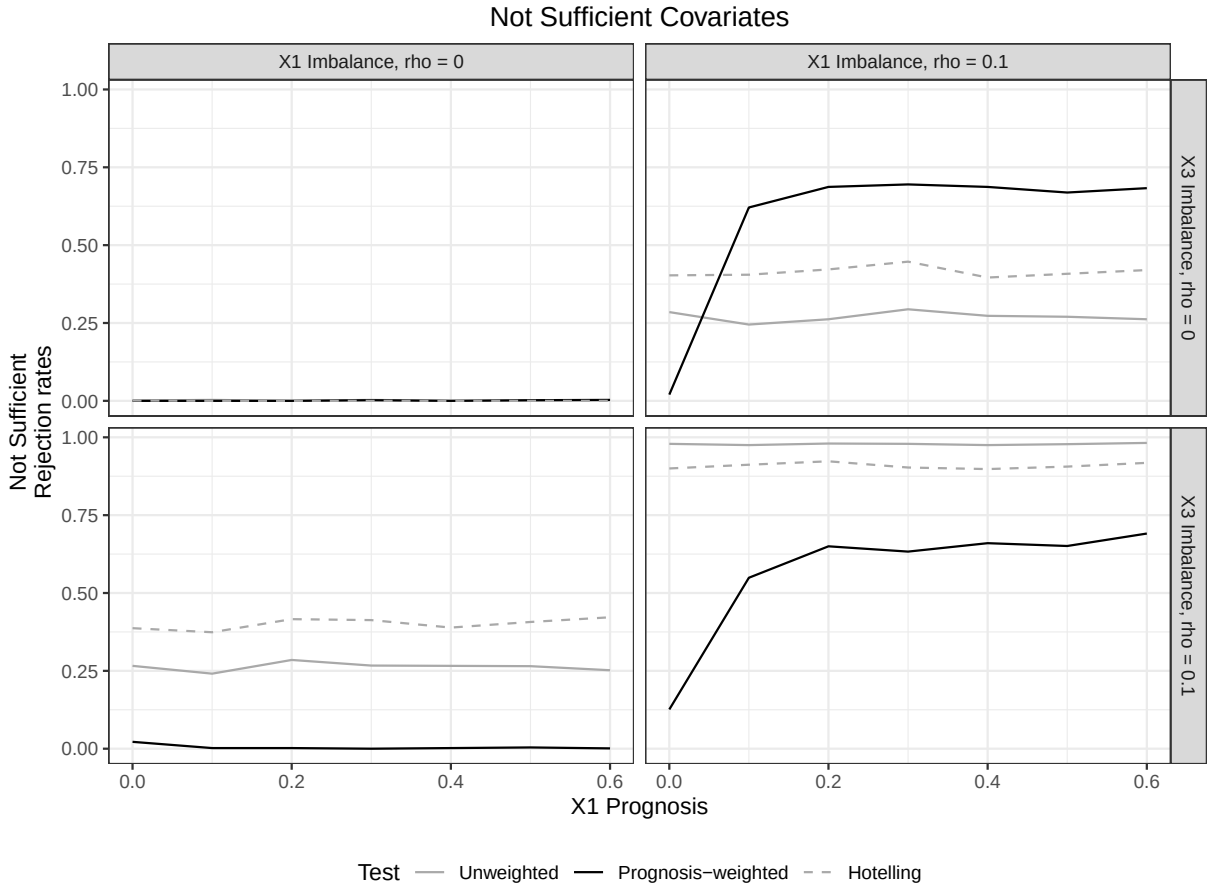


Figure 4: The simulation set up is the same as in Figure 2. In this case we consider tests that use signal covariate X_1 and a noise covariate X_3 independent of potential outcomes, therefore omitting prognostic covariate X_2. The figure displays rejection rates as X_1 prognosis increases and for cases where X_1 and X_3 are balanced and imbalanced. X_2 is set to have a fixed prognosis of 0.25 and fixed imbalance (Corr(X2, Z) = 0.15 in expectation). Because X_2 is always imbalanced and prognostic, the null hypothesis is false in all cases.

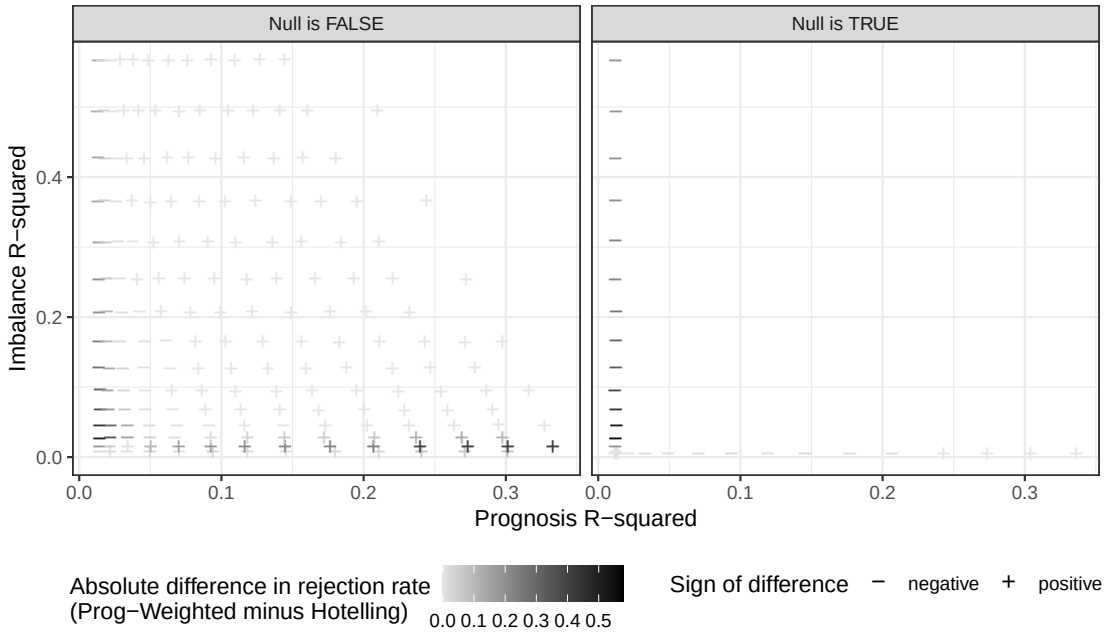


Figure 5: Difference in rejection rate of the null when it is false (left panel) and when the null is true (right panel) between the regression-weighted estimator and the unweighted estimator (Hotelling's T^2). Positive (negative) signs indicate cases where prognosis-weighted test has a higher (lower) rejection rate than Hotelling's T^2 test, with darker dots representing greater absolute difference in magnitude.

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.

Overall, these results suggest several important insights. First, the weighted and unweighted measures have similar control over Type I error when all variables are balanced. Second, however, the tests' rejection rates diverge in two important cases: when (a) as-if random is false and there is imbalance on a prognostic covariate; and (b) as-if random is true and there is imbalance on an irrelevant (non-prognostic) covariate. The results therefore suggest that the informativeness-weighted test is both more sensitive (i.e. more powerful) and more specific than unweighted tests. With moderately prognostic covariates, the weighted approach rejects with (at least weakly) higher probability when as-if random fails. And while the unweighted tests appear more powerful when covariates are only weakly prognostic, this comes at the cost of spurious rejections when as-if random holds. Indeed, the unweighted measures are completely insensitive to the degree of prognosis. Thus, they continue to reject the null with high probability due to imbalance on irrelevant covariates, even as prognosis increase. With prognostic covariates, the weighted test improves on the unweighted tests: it rejects as-if random at least as often as do unweighted tests when Assumption 1 fails while—unlike the unweighted tests—it is less prone to rejection when as-if random is true.

5.2 Application to natural experiments

Using our open-source `R` software package, we applied our prognosis-weighted and unweighted measures to the sample of natural experiments presented in Figure 1. Figure 6 reports p-values from the resampling-based hypothesis tests outlined in section 4.4.

Several points are useful. First, as we have emphasized, the prognosis-weighted measure (like the unweighted measure) provides a single metric by which to evaluate the strength against as-if random. It thus does not suffer from the indeterminacy of covariate-by-covariate tests, for which there is not typically a clear rejection rule due to dependency across covariates and multiple testing problems. Note that the permutation test in equation (10) from which the p-values are derived also accounts in a natural for the dependency across covariates, as it uses the randomization distribution of the test statistics.

Second, consistent with the sampling of ostensible natural experiments in Figure 1, we fail to reject as-if random for half of these papers. Using the prognosis-weighted measure, we reject as-if random in 6 of the 12 studies, compared to 3 such cases using the unweighted balance test. Note that using a broader sample of observational studies, we might instead fail to reject as-if random where other tests might reject it, for example, when we observe imbalance on irrelevant (non-prognostic) covariates.

Finally, while in some cases the prognosis-weighted and unweighted p-values converge, in many they diverge substantially. Intuitively, the latter case tends to occur when the prognostic value of covariates is unequally distributed, for example, when there is a single prognostic covariate that is unbalanced but other non-prognostic covariates are balanced. In Figure 7 in Online Appendix Subsection 7.9, we plot the difference of means associated with each covariate in each study against each covariate's standardized regression coefficient, as well as the overall prognosis R^2. In some studies, the prognosis of different covariates is more roughly equal, while in others one or more covariates are much more prognostic than

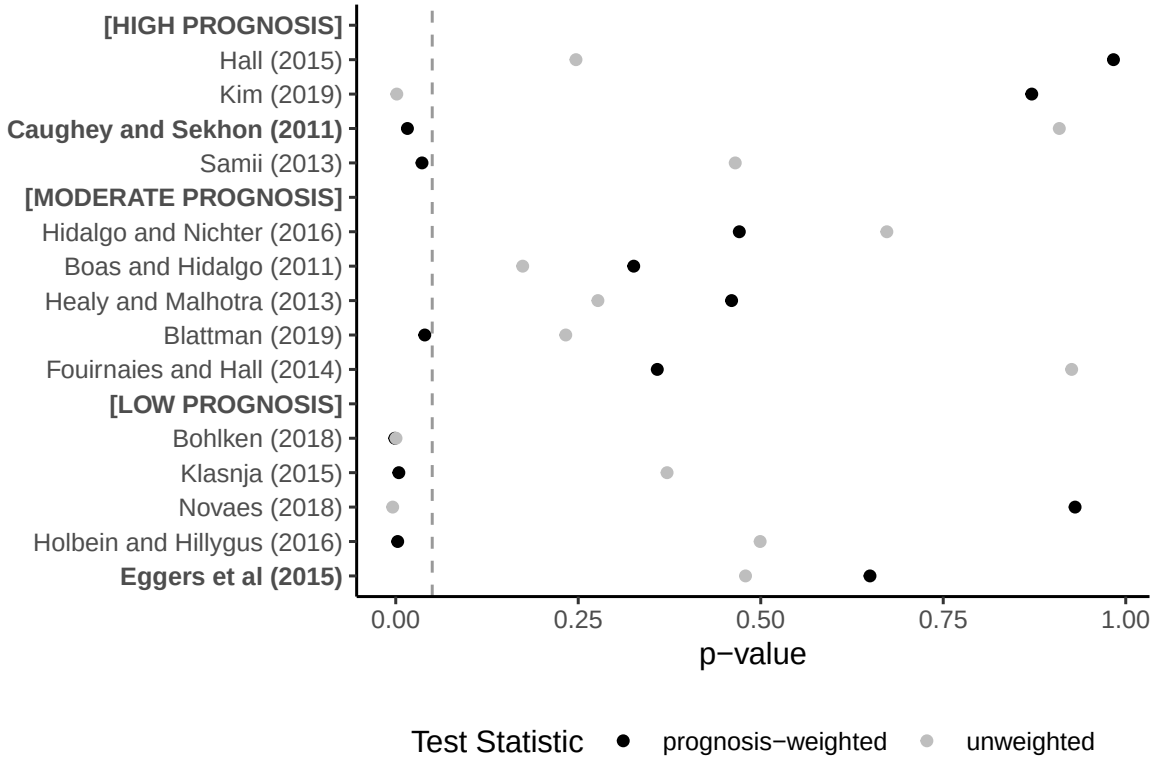


Figure 6: The figure plots p-values for prognosis-weighted and unweighted measures of covariate balance. The studies are from the sample of natural experiments in Figure 1. The dotted vertical line is at $p = 0.05$.

the others. These prognostic covariates receive more weight in the weighted test statistic δ_{PW} and thus results from the test will depend more heavily on the balance or imbalance of those key, informative covariates.

6 Discussion and conclusion

Prognosis is a critical consideration for balance testing. When covariates are not predictive of potential outcomes, balance tests may be uninformative about failures of as-if random. Treatment assignment may depend on potential outcomes, yet balance tests may fail to detect it. Conversely, imbalances on irrelevant covariates could lead to spurious rejections even when as-if random holds.

We have thus proposed that balance tests should be weighted by the prognostic power of the covariates being tested. We presented an informativeness-weighted test in which potential outcomes are first projected onto covariates. The test statistic is then a weighted sum of the individual covariate differences of means, where the weights are the standardized regression coefficients. Our theory and simulations show that this approach reduces false positive rejections at low levels of prognosis and increases power at low levels of imbalance. Covariate balance testing tends to be more powerful and more specific when the covariates as a whole are more prognostic; when covariates are minimally sufficient, the non-independence of covariates and treatment assignment implies the non-independence of treatment assignment and potential outcomes. Our procedure can be seen as an attempt to construct a minimally sufficient set of covariates, but this will work best when the covariates in question are jointly prognostic. Thus, the predictive value of the covariates for potential outcomes matters greatly for the performance of balance tests.

Our results suggest several recommendations for practice. First and perhaps most importantly, we urge researchers to present diagnostic measures of the prognostic power of the covariates in their balance tests, which they currently only rarely discuss. We recommend a simple multiple R^2 from the regression of potential outcomes under control (using control group observations) on all covariates. We prefer this to an adjusted R^2, which penalizes the R^2 as the number of covariates increases, for both theoretical and practical reasons. Theoretically, the important question is whether covariates are jointly sufficient, not how many covariates one must measure to achieve sufficiency. Practically, it is useful to incentivize researchers to gather as many covariates as they can to maximize prognosis. The use of a single omnibus test mitigates problems of multiple comparisons that might otherwise arise from the use of many covariates.

Our theoretical and simulation results suggest that tests of as-if random will be more powerful and more specific the stronger the prognosis R^2 is. Thus, we urge researchers to report it as a diagnostic statistic, along with perhaps other measures of goodness of fit. There is likely no single threshold value of this R^2 that will provide adequate protection against both false positives and false negatives that is valid across different data structures. In our simulations, the informativeness-weighted test provides good protection against both false negatives and false positives with an R^2 as low as .10, but this reflects the number of included covariates and the specifics of the simulation. Researchers and their readers should therefore take the R^2 as only one indicator of prognosis. It seems good practice, where feasible, to include a pre-treatment (lagged) measure of the outcome; our theory in Section 3 provides a rationale for doing this in terms of the sufficiency of covariates. Yet, as our motivating example in Section 2 suggests, a lagged dependent variable may not in fact be prognostic, depending on the context. More generally, the extent to which any covariate or set of covariates is in fact predictive of potential outcomes is an empirical question, which underscores the importance of diagnostic measures such as the prognosis R^2.

Second, researchers should present omnibus p-values, as in the informativeness-weighted test we have

presented here. One point of statistical testing is to provide decision rules for rejections of null hypotheses, but this is not readily allowed current standard practice in political science and other disciplines. Moreover, covariate-by-covariate balance tests do not provide a clear rejection rule that is valid in the presence of correlated covariates and dependent tests. The global p-values derived from our test, in contrast, allow for a decisive rejection rule.

Third, because unweighted tests may be more powerful at very low levels of prognosis, a conservative approach that protects against Type I error may be to present omnibus p-values from both weighted and unweighted tests. The most powerful unweighted tests in our simulations used Hotelling's T^2 statistic, so that may be a good choice for an unweighted test. However, using unweighted tests risks spurious rejections due to imbalances on non-prognostic covariates. The best way to protect against this is to use the weighted test with prognostic covariates, since this protects against both false negatives and false positives.

Finally, researchers may also wish to report separate covariate-by-covariate tests, as in often current standard practice. Like reporting the underlying components of an index, this will allow readers to continue to assess informally the overall weight of the evidence in support or against as-if random and boosts transparency, even as readers may perhaps give priority to the more powerful and specific global measure. It may also be valuable to present the difference of means for each covariate, the prognosis coefficients and prognosis and imbalance R^2s in the same table or graph. The prognosis coefficients, i.e. the standardized coefficients from the regression of potential outcomes on all covariates—or the standard deviation of those coefficients—can give a sense of the extent to which the prognosis is or is not equally distributed across covariates. This can give insight into discrepancies between weighted and unweighted tests, per our discussion of applications in Section 4.

Our approach also suggests several possible extensions for future research. We use linear regression to derive the weighted sum in equation (2). This provides a clear interpretation of the weights as measures of linear prognosis, or what we call informativeness; in addition, our main interest is in using a sample version of the linear regression to estimate the average value of potential outcomes under control in the treatment group, for which a linear regression should be adequate. However, one future extension to explore may be the use of non-linear approximations to $E(Y_i|X)$ in the finite population, such as lowess regressions. Also, we focus on the regression of $Y(0)$ rather then $Y(1)$ on covariates; one reason is that covariates such as the lagged (pre-intervention) outcome may be most prognostic for $Y(0)$, especially when treatment effects are not the same for all units (see Hansen 2008). However, future research could explore regressions that make use of information on $Y(1)$.

Our approach also has important connections to sensitivity analysis which can be further explored. Covariates that are related to potential outcomes—i.e. that have larger standardized β_j—are potentially confounders when estimating treatment effects, if they are also imbalanced across the treatment and control groups. Indeed, confounders must be associated with both a putative cause and an outcome. Existing balance tests, however, look only at the association with treatment (the putative cause) but not with the outcome. They then conclude from evidence of balance that confounding is not a problem (i.e., as-if random holds); or from evidence of imbalance that as-if random should be rejected. Yet as we have shown, this reasoning is misleading: we could find imbalance on irrelevant ("noise") covariates and yet as-if random could fail, or we could find balance on those covariates and yet treatment assignment could depend on potential outcomes. Future work could use our test in connection with sensitivity analyses that explore robustness to confounding selection biases.

We have focused here, however, on finding the most powerful and specific tests of as-if random. For

natural experiments and related designs, assessing as-if random comes prior to the estimation of treatment effects and thus of any statistical adjustment for confounders or any formal sensitivity analysis. Results from our test can certainly provide guidance regarding which covariates may need to be adjusted in estimating treatment effects, should the key identifying assumption fail. If powerful balance tests such as the one we have proposed provide evidence consistent with as-if random, however, statistical adjustment and sensitivity tests become less relevant—and the advantages of design-based analysis more pronounced.

References

Abadie, Alberto, C. M. M. and West, M. R. (2018). Endogenous stratification in randomized experiments. *Review of Economics and Statistics*, 100(4):567–580.

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

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–18.

Cattaneo, M. D., Frandsen, B. R., and Títiunik, 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.

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.

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

Eggers, A., Tuñón, G., and Dafoe, A. (2021). Placebo tests for causal inference.

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

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

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Gagnon-Bartsch, J. and Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, 13(3):1464 – 1483.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3).

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.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219 – 236.

Hartman, E. (2021). Equivalence testing for regression discontinuity designs. *Political Analysis*, 29(4):505–521.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2).

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51 – 71.

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. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190.

Leacy, F. and Stuart, E. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med*, 33(20):3488–508.

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.

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. (1988a). *Probabilistic Reasoning in Intelligent Systems*. Elsevier.

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

Rosenbaum, P. (2010). *Design of Observational Studies*.

Rosenbaum, P. R. (2002). Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, 17(3):286 – 327.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Ran- domized and Nonrandomized Studies. *Journal of Educational Psychology*, (66):688–701.

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.

Schiumerini, L. E. (2015). *Incumbency and Democracy in South America*. PhD thesis, Yale University.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12(1):487–508.

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.

Wainstein, L. (2022). Targeted Function Balancing.

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.

Westfall, P. (2005). *Combining P Values*.

Zhao, A. and Ding, P. (2021). Covariate-adjusted fisher randomization tests for the average treatment effect.

7 Online Appendix

7.1 Prognosis of lagged incumbency (Eggers et al.)

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

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

7.2 Sample of published studies

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.[43]

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.[44] 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.

7.2.1 Included Studies

Blattman (2009)

Blattman (2009) studies the relationship between violence and political participation of excombatants. The paper rests on the assumption that abduction into the Lord's Resistance Army (and consequent experience of violence) was exogenous. We focus our analysis on the first outcome studied: whether individual voted in 2005, and use the covariates the author lists the balance tests in Table 2. We restrict the analysis to interviewed subjects only, so that we the prognogis as for the 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.

Bohlken (2018)

Bohlken (2018) is interested in measuring partisan bias in the allocation of public resources. She 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

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

[44] 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.

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 her balance test.

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 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."

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 randomly pick the latter specification for 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 (this does not include number of siblings, a covariate that is included in the specifications in equations (1) and (2) of the paper).

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 'electoratechange0407.perpop' as a covariate, which is the change in electorate between 2004 and June 2007 and as a percentage of 2007 population.

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 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).

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. It is worth noting that the RD analysis actually pools outcome data across multiple years (with year fixed effects), but for the prognosis analysis we use outcome data from 1921 (the first post-treatment year observed in the data set). Our analysis includes all covariates authors test balance on in Figures SI 1.1-1.3: left parties vote share in 1917, turnout in 1917, ENEP in 1917, share of organized citizens, percentage of female attendees in municipal meetings in 1917, 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.

Klašnja (2015) Our analysis focuses on Klasnja (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).

Malesky, Nguyen, and Tran (2014)
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 0.05% (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. 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. 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.

7.2.2 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.

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).[45]

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

7.3 Proof of Theorem 1 (Sufficiency of Covariates)

To prove Theorem 1, we use alternate definitions of sufficiency and minimal sufficiency.

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

Definition 2. *Minimal sufficiency of covariates (alternate version of Definition 1)*

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

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

Lemma 1. *Definition 1 holds* $\iff$ *Definition 2 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 2, 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 1988b). Thus, Definition 1 can hold if and only if Definition 2 holds.

With these preliminaries, we can prove Theorem 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.

7.4 Proof of Theorem 2 (distribution of δ_{UW})

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 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

$$\sigma_u^2 = \frac{1}{N}\sum_i^N (u_i - \overline{u})^2 \tag{11}$$

and

$$\sigma_v^2 = \frac{1}{N}\sum_i^N (v_i - \overline{v})^2, \tag{12}$$

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

$$\sigma_{u,v} = \frac{1}{N}\sum_i^N (u_i - \overline{u})(v_i - \overline{v}). \tag{13}$$

Let $\overline{U}_z$ denote the sample average in the treatment ($z = 1$) or control ($z = 0$) group, and similarly for $\overline{V}_z$.[46] 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

$$\mathrm{Cov}(\overline{U}_1, \overline{V}_1) = \frac{N - n_1}{N - 1}\frac{\sigma_{u,v}}{n_1} = \frac{n_0}{N - 1}\frac{\sigma_{u,v}}{n_1}, \tag{14}$$

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

$$\mathrm{Cov}(\overline{U}_0, \overline{V}_0) = \frac{N - n_0}{N - 1}\frac{\sigma_{u,v}}{n_1} = \frac{n_1}{N - 1}\frac{\sigma_{u,v}}{n_1} \tag{15}$$

(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

$$\mathrm{Var}(\overline{U}_1) = \frac{N - n_1}{N - 1}\frac{\sigma_u^2}{n_1} = \frac{n_0}{N - 1}\frac{\sigma_u^2}{n_1} \tag{16}$$

and

$$\mathrm{Var}(\overline{V}_0) = \frac{N - n_0}{N - 1}\frac{\sigma_v^2}{n_0} = \frac{n_1}{N - 1}\frac{\sigma_v^2}{n_0}. \tag{17}$$

Using combinatorial calculations (see Freedman et al. 2007: A32-34 or Neyman et al. 1923), the covariance of the sample averages is

$$\mathrm{Cov}(\overline{U}_1, \overline{V}_0) = -\frac{\sigma_{u,v}}{N - 1}. \tag{18}$$

[46] 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.

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

$$\begin{aligned}
\text{Var}(\overline{U}_1 - \overline{V}_0) &= \text{Var}(\overline{U}_1) + \text{Var}(\overline{V}_0) - 2\text{Cov}(\overline{U}_1, \overline{V}_0) \\
&= \frac{1}{N-1}\left[\frac{n_0\,\sigma_u^2}{n_1} + \frac{n_1\,\sigma_v^2}{n_0} + 2\,\sigma_{u,v}\right] \\
&= \frac{1}{N-1}\left[\frac{n_0^2\,\sigma_u^2 + n_1^2\,\sigma_v^2 + 2n_1(n_0)\,\sigma_{u,v}}{n_1(n_0)}\right], \qquad (19)
\end{aligned}$$

using (16), (17), and (18) 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.[47]

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

$$\begin{aligned}
\text{Var}(\delta_{UW}) &= \text{Var}(\sum_{j=1}^{p} \delta_j) \\
&= \sum_{j=1}^{p} \text{Var}(\delta_j) + 2\sum_{j<k}^{p} \text{Cov}(\delta_j, \delta_k). \qquad (20)
\end{aligned}$$

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 (19), we find

$$\begin{aligned}
\text{Var}(\delta_j) &= \text{Var}(\overline{X}_{j1} - \overline{X}_{j0}) \\
&= \frac{1}{N-1}\left[\frac{n_1\,\sigma_{X_j}^2}{n_0} + \frac{(n_0)\,\sigma_{X_j}^2}{n_1} + 2\,\sigma_{X_j}^2\right] \\
&= \frac{1}{N-1}\left[\frac{n_1^2\,\sigma_{X_j}^2 + n_0^2\sigma_{X_j}^2 + 2n_0(n_1)\,\sigma_{X_j}^2}{n_0(n_1)}\right] \\
&= \frac{1}{N-1}\left[\frac{(n_0+n_1)^2\sigma_{X_j}^2}{n_0(n_1)}\right]. \\
&= \frac{1}{N-1}\left[\frac{N^2\sigma_{X_j}^2}{n_0(n_1)}\right]. \qquad (21)
\end{aligned}$$

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.[48] Also, $\sigma_{X_j}^2$ is (11) with $u_i = X_{ij}$: it denotes the variance of the covariate X_j calculated

[47] 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.

[48] 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 vector Z.

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

$$\sigma^2_{X_j} = \frac{1}{N}\sum_{i=1}^{N}(X_{ij} - \overline{X}_j)^2, \tag{22}$$

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.[49] 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 (21) or (30) 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

$$\begin{aligned}\text{Cov}(\delta_j, \delta_k) &= \text{Cov}(\overline{X}_{j1} - \overline{X}_{j0}, \overline{X}_{k1} - \overline{X}_{k0}) \\ &= \text{Cov}(\overline{X}_{j1}, \overline{X}_{k1}) - \text{Cov}(\overline{X}_{j1}, \overline{X}_{k0}) - \text{Cov}(\overline{X}_{j0}, \overline{X}_{k1}) + \text{Cov}(\overline{X}_{j0}, \overline{X}_{k0}).\end{aligned} \tag{23}$$

The first and fourth terms in (23) are the covariances of the sample averages of two features, both sampled without replacement from a finite population of size N. Using (14) and (15), we have

$$\text{Cov}(\overline{X}_{j1}, \overline{X}_{k1}) = \frac{N - n_1}{N - 1}\frac{\sigma_{X_j,X_k}}{n_1} = \frac{n_0}{N - 1}\frac{\sigma_{X_j,X_k}}{n_1} \tag{24}$$

and

$$\text{Cov}(\overline{X}_{j0}, \overline{X}_{k0}) = \frac{N - n_0}{N - 1}\frac{\sigma_{X_j,X_k}}{n_0} = \frac{n_1}{N - 1}\frac{\sigma_{X_j,X_k}}{n_0}. \tag{25}$$

The second and third terms in (23) 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 (18), we have

$$\text{Cov}(\overline{X}_{j1}, \overline{X}_{k0}) = \text{Cov}(\overline{X}_{j0}, \overline{X}_{k1}) = -\frac{1}{N - 1}\sigma_{X_j,X_k}, \tag{26}$$

where σ_{X_j,X_k} is the population covariance given in (13), with the covariates X_j and X_k in place of u and v.

[49] 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,

$$\begin{aligned}
\mathrm{Cov}(\delta_j, \delta_k) &= \frac{n_0}{N-1}\frac{\sigma_{X_j,X_k}}{n_1} + 2\frac{\sigma_{X_j,X_k}}{N-1} + \frac{n_0}{N-1}\frac{\sigma_{X_j,X_k}}{n_0} \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0}{n_1} + 2 + \frac{n_1}{n_0}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0^2}{n_1(n_0)} + \frac{2n_1(n_0)}{n_1(n_0)} + \frac{n_1^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{n_0^2 + 2n_1(n_0) + n_1^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{(n_0+n_1)^2}{n_1(n_0)}] \\
&= \frac{\sigma_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}].
\end{aligned} \tag{27}$$

Returning to (20) and substituting for $\mathrm{Var}(\delta_j)$ and $\mathrm{Cov}(\delta_j, \delta_k)$, we have

$$\begin{aligned}
\mathrm{Var}(\delta_{UW}) &= \mathrm{Var}(\sum_{j=1}^{p}\delta_j) \\
&= [\sum_{j=1}^{p}\mathrm{Var}(\delta_j) + 2\sum_{j<k}^{p}\mathrm{Cov}(\delta_j, \delta_k)] \\
&= [\sum_{j=1}^{p}\frac{1}{N-1}\left[\frac{N^2\sigma_{X_j}^2}{n_0(n_1)}\right] + 2\sum_{j<k}^{p}\frac{\sigma_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}] \\
&= [\frac{N^2}{N-1}\left[\frac{1}{n_0(n_1)}\sum_{j=1}^{p}\sigma_{X_j}^2\right] + 2\frac{N^2}{N-1}\frac{1}{n_1(n_0)}\sum_{j<k}^{p}\sigma_{X_j,X_k}] \\
&= \frac{N^2}{N-1}\frac{1}{n_0(n_1)}\left[\sum_{j=1}^{p}\sigma_{X_j}^2 + 2\sum_{j<k}^{p}\sigma_{X_j,X_k}\right].
\end{aligned} \tag{28}$$

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.[50]

Note that when the covariate X_j is standardized as

$$(X_{ij} - \overline{X_j})/\sigma_j, \tag{29}$$

[50]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 (21) that

$$\text{Var}(\delta_{j,\text{stand}}) = \frac{N^2}{N-1}\left[\frac{1}{n_0(n_1)}\right], \tag{30}$$

and σ_{X_j,X_k} in (27) is ρ_{X_j,X_k}, the correlation of $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$. Then

$$\text{Var}(\delta)_{UW,\text{stand}}) = \frac{N^2}{N-1}\frac{1}{n_0(n_1)}\left[p + 2\sum_{j<k}^{p}\rho_{X_j,X_k}\right]. \tag{31}$$

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.

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

Hotelling's T^2 statistic is directly related to δ_{UW}, the unweighted sum of covariate differences of means, as follows. 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 (point 3. of the proof of Theorem 2 in subsection 7.4). 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 section 5, but they are all unresponsive to

the level or distribution of covariate prognosis.

7.6 Proof of Theorem 3 (conditional distribution of δ_{PW})

Consider the distribution of

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

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.

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

As for the conditional variance of δ_{PW},

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

where in the final line we use (20) and (27). When the elements of $\boldsymbol{X}$ are standardized, we have

$$\text{Var}(\delta_{RW,\text{stand}}|\widehat{\beta}) = \sum_{j=1}^{p} \widehat{\beta_j^C}^2 \frac{1}{N-1}\left[\frac{N^2}{n_0(n_1)}\right] + 2\sum_{j<k}^{p} \widehat{\beta_j^C}\widehat{\beta_k^C}\frac{\rho_{X_j,X_k}}{N-1}[\frac{N^2}{n_1(n_0)}] \tag{32}$$

In sum, the conditional variance of δ_{PW} 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 tend to vanish.

7.7 Proof of Theorem 4 (Informativeness-Weighted Covariate Balance Test)

The statement of Theorem 4 (Observable Implications of As-If Randomization): Suppose that $\boldsymbol{X}$ is sufficient for $\{\boldsymbol{Y(1)}, \boldsymbol{Y(0)}\}$. Then $\boldsymbol{Z} \not\perp\!\!\!\perp \widehat{\boldsymbol{Y}(\boldsymbol{Z})}|\boldsymbol{X} \implies \boldsymbol{Z} \not\perp\!\!\!\perp \boldsymbol{Y}(\boldsymbol{Z})$

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 (3) for $Y(0)$; randomness in $\widehat{\beta_j^Z}$ is induced by treatment assignment (see e.g. footnote 28). 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, $\forall U \neq X, Y(Z) \perp\!\!\!\perp U|\boldsymbol{X}$. 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

$$\widehat{\gamma} = \frac{\widehat{\text{Cov}}(\boldsymbol{Y(0)}, \widehat{\widetilde{\boldsymbol{U}}})}{\widehat{\text{Var}}(\widehat{\widetilde{\boldsymbol{U}}})}, \tag{33}$$

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 3) and sufficiency, $\text{E}(\widehat{\gamma}) \to 0$, and

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

for arbitrarily small ϵ.

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

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

7.8 Simulation Results

Table A3: Simulated rejection rates for each test when covariates are minimally sufficient

progX1	imbalX1	progX2	imbalX2	UW p	PW p	Hotelling p	Rsqr prog	Rsqr bal
0	0	0.25	0	0.004	0.001	0	1	0.002
0.25	0	0.25	0	0.001	0	0	1	0.001
0.5	0	0.25	0	0.003	0.003	0	1	0.002
0	0.05	0.25	0	0.031	0.002	0.025	1	0.004
0.25	0.05	0.25	0	0.03	0.027	0.019	1	0.004
0.5	0.05	0.25	0	0.026	0.055	0.016	1	0.004
0	0.1	0.25	0	0.277	0.001	0.433	1	0.012
0.25	0.1	0.25	0	0.273	0.276	0.425	1	0.012
0.5	0.1	0.25	0	0.267	0.535	0.419	1	0.011

Note: Signal covariates are X1 and X2 and covariates included in the global tests are X1 and X2. Covariate X2 is balanced in expectation and has a fixed prognosis parameter value (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 Section 5.

Table A4: Simulated rejection rates for each test when covariates are sufficient

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

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 Section 5.

Table A5: Simulated rejection rates for each test when covariates are not sufficient

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

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 5.

7.9 The distribution of prognosis in sampled natural experiments

For each of the natural experimental studies included in Figures 1 and Figure 6, we plot in Figure 7 the standardized difference of means on each covariate against that covariate's standardized prognosis regression coefficient (i.e. we plot each δ_j against each $\widehat{\beta_j^C}$ in equation 2).

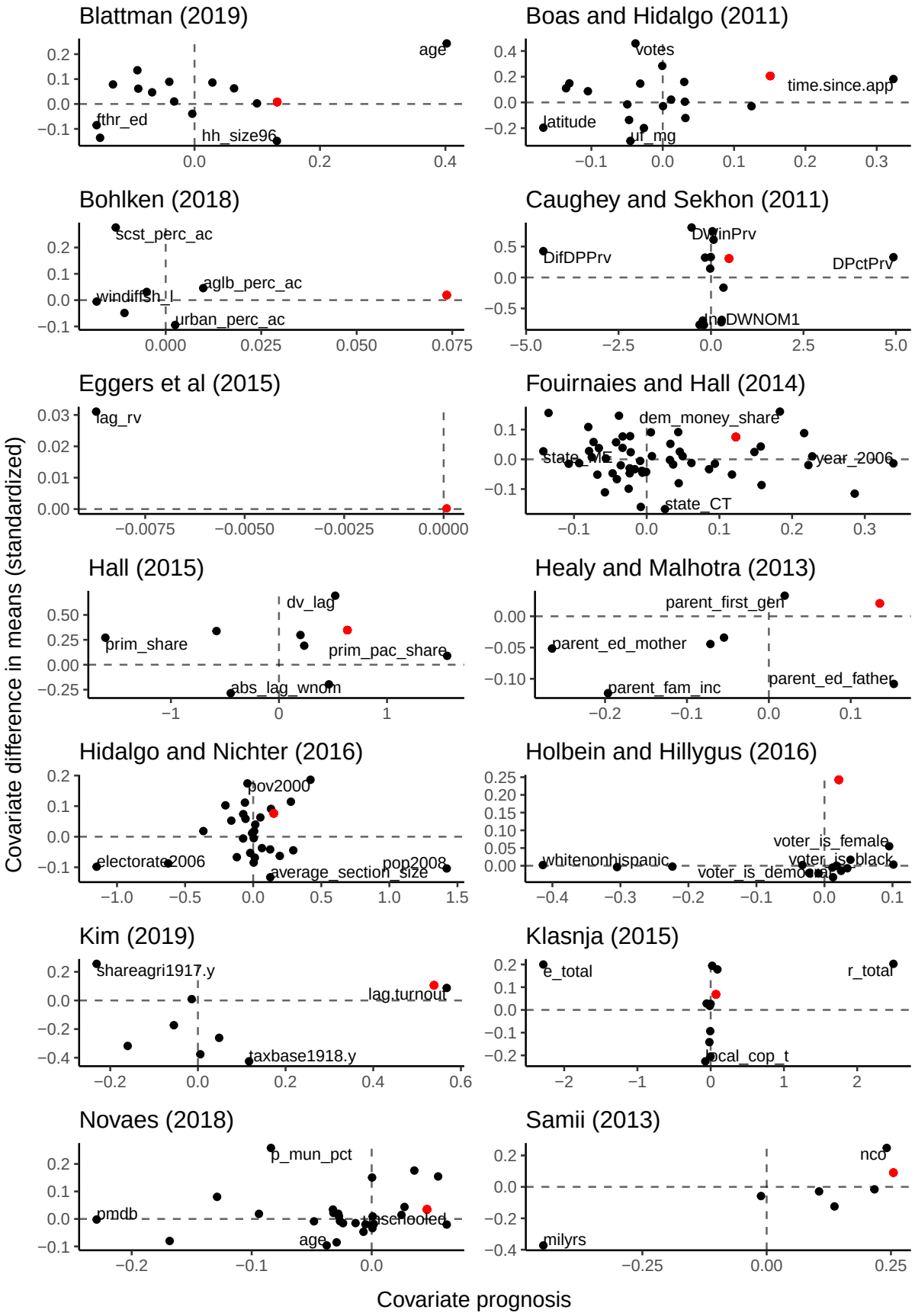


Figure 7: For each of the natural experimental studies included in Figure 6, we plot the standardized difference of means on each covariate against that covariate's standardized prognosis regression coefficient. The red dots indicate the overall prognosis and imbalance R^2s.