HHS Public Access

Author manuscript

Am Stat. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

Am Stat. 2023 ; 77(1): 72–84. doi:10.1080/00031305.2022.2051605.

Assignment-Control Plots: A Visual Companion for Causal Inference Study Design

Rachael C. Aikens[*],

Department of Biomedical Data Science, Stanford University

Michael Baiocchi

Department of Epidemiology and Population Health, Stanford University

Abstract

An important step for any causal inference study design is understanding the distribution of the subjects in terms of measured baseline covariates. However, not all baseline variation is equally important. We propose a set of visualizations that reduce the space of measured covariates into two components of baseline variation important to the design of an observational causal inference study: a propensity score summarizing baseline variation associated with treatment assignment, and prognostic score summarizing baseline variation associated with the untreated potential outcome. These *assignment-control plots* and variations thereof visualize study design trade-offs and illustrate core methodological concepts in causal inference. As a practical demonstration, we apply assignment-control plots to a hypothetical study of cardiothoracic surgery. To demonstrate how these plots can be used to illustrate nuanced concepts, we use them to visualize unmeasured confounding and to consider the relationship between propensity scores and instrumental variables. While the family of visualization tools for studies of causality is relatively sparse, simple visual tools can be an asset to education, application, and methods development.

Keywords

propensity score; prognostic score; matching; instrumental variable; observational study; visualization

1 INTRODUCTION

In an observational study, individuals “self-select” into treated and untreated groups, rather than being randomly assigned by an experimenter. Thus, in order to make claims about causal effects, the researcher must address the possibility of bias due to systematic differences in the baseline characteristics of treated and untreated individuals. A host

[*]Rachael C. Aikens is a PhD Candidate, Stanford University Department of Biomedical Data Science, Stanford, CA, 94305 (raikens@stanford.edu).

Michael Baiocchi is an assistant professor of Epidemiology and Population Health at Stanford University, Stanford, CA, 94305. M. Baiocchi also has affiliations with the Stanford University Department of Statistics and the Stanford University Department of Biomedical Data Science.

Supplementary Mathematics Notes

A brief proof for the claim that $\rho = Corr(\phi(X), \Psi(X))$ in the main simulation set up from Section 2.2. (.pdf)

of matching and conditioning methods have been proposed to do just this (Stuart 2010, Rosenbaum 2020, Imbens and Rubin 2015). A question fundamental to these approaches is: How should the compared treated and untreated samples be similar (or different) in order to obtain a clear understanding of the causal effect? What different types of baseline variation – in general – are important to a causal inference study?

One popular candidate is the propensity score. Many subclassification or adjustment procedures apply an estimated propensity score, which summarizes the measured baseline variation influencing the probability of assignment. Intuitively, propensity score methods model the treatment assignment mechanism based on observed covariates, so that it can be adjusted for. Under suitable assumptions (including no unmeasured confounding), matching exactly on the propensity score recapitulates a completely randomized controlled experiment, allowing for identification and unbiased estimation of treatment effect. However, critics of propensity score matching note that the propensity score tends to neglect baseline variation that is less associated with treatment assignment but influential on the *potential outcomes* of the study subjects, potentially resulting in unfavorably high variance and low statistical power (King and Nielsen 2019).

The less-commonly discussed prognostic score, formalized by Hansen (2008), models the expected outcome of each subject in the absence of treatment, based on the observed covariates. Interestingly, under suitable assumptions, balancing on the prognostic score results in a form of covariate balance that leads to unbiased estimation of the causal effect, analogous to the propensity score (Hansen 2008). Balancing on a prognostic score may also convey additional statistical benefits: increasing precision and yielding results less easily explained away by unobserved confounding (Rosenbaum 2005 a, Aikens et al. 2020). A small but growing body of literature suggests that matching methods which apply both a prognostic score and a propensity score may be a favorable approach in some observational contexts (Leacy and Stuart 2014, Antonelli et al. 2018, Aikens et al. 2020).

Assignment-control plots (AC plots), introduced briefly by Aikens et al. (2020), visualize each subject in an observational data set in terms of their propensity and prognostic score. These are two (often interrelated) summaries of baseline variation that are directly relevant to observational studies of causality: propensity score similarity between compared individuals reduces bias (Rosenbaum and Rubin 1983), while prognostic score similarity between compared individuals reduces bias as well as variance and increases power in sensitivity analyses of unobserved confounding (Hansen 2004, Leacy and Stuart 2014, Antonelli et al. 2018, Aikens et al. 2020). Akin to a dimensionality reduction, propensity and prognostic scores can digest a potentially complex data set – real or simulated – into distinct components which are statistically and practically meaningful. AC plots are scatter plots visualizing how these components of baseline variation are jointly distributed in a data set.

Assignment-control plots are both a multipurpose diagnostic in application and a versatile conceptual illustration for methods development and education. In this paper, we showcase several examples of assignment-control plots, ranging from applied diagnostic visualizations (Sections 3.1, 3.2) to conceptual illustrations of primary interest to educators and causal

inference methodologists (Sections 3.3 and 3.4). From an applied perspective, clear visualizations can guide conversations on design trade-offs with non-technical collaborators, either as an early data diagnostic (Section 3.1) or while comparing matching schemes (Section 3.2). In the latter half of this paper, we showcase how AC plots can help frame relevant topics in methodological research, including unobserved confounding (Section 3.3), and the relationship between propensity score and instrumental variable approaches (Section 3.4). By visualizing core causal inference concepts, each of these sections also suggest the educational potential of AC plots. Most examples in this report are illustrations from simulated data, and we conclude with an example of assignment-control diagnostic plots in a hypothetical study of cardiothoracic surgery.

2 METHODS

2.1 Notation and background

We adopt the Neyman-Rubin potential outcomes framework, in which a sample is described by

$$\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^{n},$$

where the triplet (X_i, T_i, Y_i) describes an individual with measured covariates X_i, binary treatment assignment indicator T_i, and observed outcome Y_i. Often, X_i is a vector of many characteristics. We take $Y_i(t)$ to represent the potential outcome of individual i under a specific treatment assignment, t. The fundamental problem of causal inference is that it is impossible to observe both potential outcomes, $Y_i(0)$ and $Y_i(1)$ for any individual.

The propensity score is defined as $e(X) = P(T = 1|X)$. The popularity of the propensity score in observational studies stems primarily from its use as a balancing score, i.e.

$$T \perp X \mid e(X) \tag{1}$$

That is, within level-sets of the propensity score, the treatment assignment is independent of the measured covariates. Under the assumption of strongly ignorable treatment assignment, exact matching on the propensity score allows for unbiased estimation of the treatment effect (Rosenbaum and Rubin 1983).

The prognostic score is defined by Hansen as any quantity $\Psi(X)$ such that

$$Y(0) \perp X \mid \Psi(X) \tag{2}$$

In essence, a prognostic score is any function of the measured covariates that – through conditioning – induces independence between the potential outcome in the absence of treatment ($Y(0)$) and the measured covariates. It is thus, by definition, a balancing score as well. Under regularity conditions analogous to those for the propensity score, conditioning on the prognostic score also allows for unbiased estimation of the treatment effect, as described in more detail by Hansen (2008). When $Y(0)|X$ follows a generalized linear model

$\Psi(X) = E[Y(0) \mid X]$. In the literature, the prognostic score is often treated more informally as the expected outcome in the absence of treatment given the observed covariates.

2.2 Set-up

The demonstrations that follow depict several simulated data sets. In keeping with Aikens et al. (2020), the primary generative model for these is as follows:

$$X_i \sim Normal(0, I_{10})$$

$$T_i \sim Bernoulli\left(\frac{1}{1 + exp(-\phi(X_i))}\right)$$

$$Y_i(0) = \Psi(X_i) + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2),$$

where $\phi(X)$ and $\Psi(X)$ represent the true propensity and prognostic score functions. In general, these will be given by:

$$\phi(X_i) = c_1 X_{i1} - c_0$$

$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2},$$

where c_1, c_0, σ^2 and ρ are constants. In particular, the form for the prognostic function above guarantees that $\rho = Corr(\phi(X), \Psi(X))$ (see the supplementary mathematics notes for a proof). Note that X is a vector of 10 independent, normally distributed random variables, representing 10 baseline covariates measured for each individual. However, only the first two components of X_i (denoted X_{i1} and X_{i2}) are important to either the outcome or the treatment assignment. We will also briefly suggest some other possible assignment-control plots generated using different forms for ϕ and Ψ (quadratic and discontinuous, figure 2). The code for this project is available on github at https://github.com/raikens1/RACplots.

2.3 Fitting the score models

In observational studies in practice, the propensity and prognostic score models are not known. Conventionally, the propensity scores are often estimated from a logistic regression of the baseline covariates on treatment assignment, fit on the entire study sample. Fitting the prognostic score may be somewhat more nuanced (Aikens et al. 2020). First, since the prognostic model is meant to predict the outcome in the absence of treatment, the prognostic model is fit only on untreated individuals. Thus, all prognostic score estimates on the treated

individuals are necessarily extrapolations. Second, fitting the prognostic model on the entire untreated sample raises concerns of overfitting (Hansen 2004, Abadie and Imbens 2006, Antonelli et al. 2018). To address these concerns and preserve the separation of the design and analysis phases of the study, Aikens et al. (2020), propose a *pilot design,* in which a subset of the untreated individuals is selected and held aside for the purpose of fitting the prognostic model. These untreated individuals – comprising a *pilot data set* are then discarded, so that the observational units used to train the prognostic model are disjoint from the set used in the final analysis (the *analysis set*). The question of how to appropriately select the untreated observations for the pilot set is a difficult one, described in more detail elsewhere (Aikens et al. 2020, 2021).

For conceptual clarity, the theoretical examples that follow bypass the problem of score estimation by using the ground-truth propensity and prognostic scores, as specified by our simulation set-up. In section 3.5, we consider an applied example in cardiothoracic surgery in which the true propensity and prognostic score models are not known and must be estimated in order to create an assignment-control plot. For detailed discussion on the realities of fitting the score models, see Aikens et al. (2020).

2.4 Applied example

As a demonstration of assignment-control plots in practice, we consider an applied example comparing 30 day mortality between female coronary artery bypass grafting (CABG) patients with and without a female primary surgeon. This work is included as a conceptual example; a more thorough consideration of this question might include more nuanced corrections and methods not considered here. Patient covariates and outcomes were extracted from medicare claims data for 1,155,903 CABG surgeries from 1998 to 2016. The gender of the primary surgeon was obtained from the National Plan and Provider Enumeration System records. When more than one surgeon was involved in a procedure, the primary surgeon was considered to be the one with the highest volume of prior surgeries. One limitation of this study design was that patient information included only “sex ” and provider information included only “gender.” In addition, many providers had missing gender information. To protect patient privacy, random noise was added to patient age and surgery year, and medicare qualification status, race, admission type, and admission day were shuffled within groups of patients with the same outcome and exposure.

382,688 surgeries were performed on patients whose sex was recorded as female. After excluding surgeries with missing outcome information (17 observations) or missing gender information for the primary surgeon (81,233 observations), a total of 301,438 surgeries remained. We fit a logistic propensity score model from the entire data set of 301,438 surgeries. The prognostic model was fit using a logistic lasso on a pilot set of 5% of the untreated patients (nonfemale primary surgeon, 14,726 observations), leaving an analysis set of 286,712 surgeries for the remainder of the analysis. This size of a pilot set is probably unnecessarily large for most practical studies but facilitates straightforward fitting of the prognostic score for demonstration. In particular, since the outcome is known to be quite rare (30-day mortality for CABG is less than 5%)(Hansen et al. 2015), fitting the prognostic score effectively for this particular outcome is a somewhat difficult task. A more formal

consideration of the study question might consider more thorough curation and imputation of covariate information and more sophisticated modeling techniques for addressing the large covariate space. In the supplement, we include an additional example in which both male and female patients were considered and the treatment of interest was considered to be sex-gender concordance between patient and primary surgeon. All aspects of study design are the same, except that all 908,158 surgeries on male or female patients with recorded surgery date and primary surgeon gender are used (Supplementary Figure S2).

Love plots were constructed using the cobalt package (v 4.3.1) (Greifer 2021), and the pilot set was extracted using the stratamatch package (v0.1.7) (Aikens et al. 2020). The lasso prognostic model was fit using the glmnet package (v 4.0) (Friedman et al. 2010). Matches were created with DOS2 (v 0.5.2)(Rosenbaum 2019) and the optmatch package (v0.9-13) (Hansen and Klopfer 2006), which uses work from Bertsekas and Tseng (1988).

3 RESULTS

3.1 In application: Assignment-control plots and design diagnostics

At the onset of a study, it is relatively commonplace to consider the distribution of a propensity score among observational study subjects. Researchers are often cautioned to check that the treated and untreated groups overlap in propensity score and that no individuals have propensity scores of approximately 0 or 1, since these conditions are necessary (though not sufficient) to ensure that the treatment effect is identifiable and that some aspect of the treatment assignment is random Zhu et al. (2021), Petersen et al. (2012). Less attention is often paid to the distribution of the prognostic score and the relationship between propensity and prognosis in the data set. However, the connection between the likelihood of treatment and the expected untreated outcome can contain important information not captured in examinations of the propensity score alone. Informally, the relationship between propensity and prognostic scores might be said to characterize the overall nature of confounding of treatment and outcomes arising from the *measured* covariates. For example, in a medical setting where the outcome of interest is mortality (and high prognostic score means higher risk), a tight association between high propensity scores and high prognostic scores might suggest that the sickest patients are more likely to be treated.

Figure 1 shows example assignment-control plots (A-C) and propensity score density histograms (D-F) for three different simulated observational data sets from the setting in section 2.2 with different underlying levels of correlation between the true propensity and prognostic scores (ρ =0, 0.9, and −0.9). Notably, the marginal distribution of the propensity score is identical across the three simulated scenarios, and the corresponding propensity score histograms (D-F) are qualitatively equivalent. However, panels A-C reveal that the three settings are in fact quite different. In setting A, treated individuals are no different from untreated individuals in terms of their prognosis (Here, $\rho = 0$, so none of the measured covariates are actually confounders). This means that even a naive comparison of treated and untreated subjects may give an unbiased treatment effect estimate under suitable conditions. In settings B and C, propensity and prognostic scores are highly correlated (the measured covariate, X_1 is a confounder), suggesting that such a naive comparison might be highly

biased. Moreover, the direction of the correlation between propensity and prognosis suggests the direction of the bias from a naive comparison: If the individuals with the *worst* prognosis are the *least* likely to be treated (Figure 1B), we are more likely to *overestimate* the effectiveness of the treatment in producing a more positive outcome. If the individuals with the *worst* prognosis are the *most* likely to be treated (Figure 1C), we are liable to err in the *opposite* direction. Future work might consider whether this correlation indicates a tendency toward bias not only in naive comparisons, but for adjustment or subclassification approaches in which the score models or matches are imperfect.

Figure 1 also evokes a discussion of the generalizability of the estimated treatment effect to different populations. For example, many matching studies focus on estimating the sample average treatment effect among the treated individuals. However, in scenarios B and C, the treated and untreated groups are systematically quite different in terms of their prognostic score. A researcher in this position should be prepared to ask:"Can a treatment effect estimated among the healthiest individuals in my sample generalize to the sickest individuals in a population?"and vise-versa. Likewise, researchers seeking to estimate a sample average treatment effect or conditional average treatment effect might question whether such estimands can really be understood in settings in which treatment and prognosis are tightly linked (e.g. strongly correlated), and overlap between the treated and untreated samples is sparse. If, for example, few to none of the sampled treated individuals are among the sickest members of the population, this calls into question whether *any* method – however sophisticated – can confidently estimate a treatment effect that applies to this group without additional assumptions (Tipton 2014).

Table 1 lists several possible characteristics of an assignment-control plot and summarizes how these might inform the design or interpretation of a study. Since these visualizations are semi-exploratory, they can help researchers identify anomalies they might not have otherwise anticipated. For example, propensity and prognostic scores in the data may have a nonlinear relationship (Supplementary Figure S1A). This may occur in medical settings, for example, where standard treatment procedures change for patients who are the most ill. These scenarios suggest that there may be subgroups of individuals for whom treatment considerations differ, demanding especially careful consideration of treatment effect heterogeneity and generalizability. Alternatively, there may be strong sub-grouping in the data, owing to some discrete covariate (e.g., sex, recorded race, smoking status) which is strongly associated with the expected potential outcome in the absence of treatment (Supplementary Figure S1B) or the treatment assignment (Supplementary Figure S2, section 3.5). In scenarios like these, a researcher might consider stratifying or matching exactly within these groups. They might also consider whether this discrete covariate could be an important treatment effect modifier, or whether the mechanisms for prognosis and propensity are so different between these groups that they should be considered entirely separate samples with separate score models or even separate study designs.

Table 1 is not an exhaustive list. One of the greatest strengths of a scatter plot diagnostic is its versatility: myriad other possible joint distributions of propensity and prognostic scores may arise in practice, each with ramifications for study design and generalizability. For example, AC plots may also help characterize outliers in propensity or prognosis

(Supplementary Figure S3, section 3.5). Note that in *many* cases, a problem identified in an assignment control plot may not have a single agreed-upon solution or interpretation (for example, there are many alternative approaches for addressing poor overlap (Zhu et al. 2021, Petersen et al. 2012)). Where AC plots help is by allowing the researcher to diagnose problems and communicate with stakeholders in order to decide their own next steps.

3.2 In application: Assignment-control plots and matching

As described in the introduction, matches which are close in terms of the propensity score have different statistical properties than matches which are close in terms of a prognostic score. Thus, assignment control plots can be used to visualize the trade-offs between matching approaches (Figure 2). Prior work suggests that jointly applying the propensity and prognostic score to matching studies in suitable scenarios may reduce variance and increase power in gamma sensitivity analyses while increasing robustness in the case that one of the models is mis-specified (i.e. enabling doubly-robust estimation) (Aikens et al. 2020, Antonelli et al. 2018, Leacy and Stuart 2014). These findings suggest that a desirable quality for an observational study design is that matched pairs are close together in assignment-control space.

In Mahalanobis distance matching, all covariates are weighted equally in a statistical sense. When there is an abundance of uninformative covariates, Mahalanobis distance matching can select matches that may actually be quite distant in the assignment-control space (Figure 2A) (Aikens et al. 2020). On the other hand, propensity score matching optimizes directly for matches that are nearby in terms of the variation associated with the treatment (the "assignment" axis), but it is entirely agnostic to variation associated with the outcome (Figure 2B). This can result in high variance in estimated treatment effect (King and Nielsen 2019).

One interesting approach, investigated by Leacy and Stuart (2014), is to match treated and untreated individuals based on Mahalanobis distance on the full covariate space, while enforcing calipers on the prognostic and propensity scores. The two caliper methods in Figure 2 impose constraints on the matching process to ensure that matches are close in terms of propensity score (Figure 2C) or both propensity and prognostic score (Figure 2D). Figure 2D underscores why comparing prognostically similar treated and untreated individuals can make a study's results less easily explained away in gamma sensitivity analyses for unobserved confounding. Intuitively, gamma sensitivity analyses imagines some "unobserved confounding" adversary who, with a strength of Γ, shifts the treatment probabilities of matched individuals in order to bias our results. If our matched individuals are very close in terms of their likely outcomes, such an adversary can do less harm.

Assignment-control plots also visualize how matching quality can degrade when propensity score overlap between treated and untreated individuals is poor (Figure 3). If there are some levels of the propensity score at which there are many treated individuals and few untreated ones, the matching algorithm may have to reach very far away in assignment-control space in order to find adequate matches for the treated observations (Figure 3B). When this happens, the matched untreated individual is likely to be not only distant in propensity score but systematically different in prognostic score, especially when propensity and prognosis

are highly correlated. This systematic deviation can lead to bias in the effect estimate. Researchers observing AC plots like Figure 3B should consider alternatives which mitigate against poor overlap (Zhu et al. 2021, Petersen et al. 2012).

3.3 Illustration: Assignment-control plots and unmeasured confounding

The examples in this and the following section have a methodological focus, using assignment-control plots to illustrate key causal inference concepts, beginning with the problem of unmeasured confounding. A wide and increasing variety of causal inference methods for observational studies – in particular propensity score approaches – depend on the absence of unmeasured confounding. For matching studies, this dependence can be visually illustrated with assignment-control plots from simulated data.

Figure 4 illustrates the behavior of two pair-matching approaches in a scenario with unobserved confounding. We add to our data-generating set-up an unobserved confounder, U, such that:

$$\phi(X_i) = c_1 X_{i1} + \eta U - c_0,$$

$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2} + \eta U,$$

Where η is a constant determining the strength of U. Suppose the researcher somehow ascertained exactly the correct relationships between the two score models and the *observed* covariates, so that their propensity and prognostic models are precisely $\hat{\phi}(X_i) = c_1 X_{i1} - c_0$ and $\widehat{\Psi}(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2}$, respectively. That is, the score models are exactly correct, except that they do not include the unobserved confounder. Figure 4 panels A-B depict the assignment-control plots this researcher might make and the matchings they might produce using these score models, $\hat{\phi}$ and $\widehat{\Psi}$. Since both the assignment-control plots and the matchings use only the observed covariates and exclude unmeasured confounding, propensity matches appear quite close in $\hat{\phi}$ (Figure 4A) and a Mahalanobis distance matching on the informative measured covariates (X_1 and X_2) appear quite close in the assignment-control space defined by $\hat{\phi} \times \widehat{\Psi}$ (Figure 4B).

However, panels C-D in Figure 4 show the same matches in the *true* assignment-control space, in which ϕ and Ψ are known to depend on the unobserved confounder, U. In each matching, pairs tend to differ from each other due to baseline variations in the unobserved confounder that were not accounted for in the matching process. The contrast between Figures 4B and 4D most cleanly illustrate how failing to account for U results in systematic error: in the true assignment-control space, one matched individual in each pair tends to have both higher prognostic score and higher propensity score than its partner. Since this individual is more often the treated individual than the untreated individual, estimates of treatment effect based on this matching will be biased (See Supplementary Figure S5A for a schematic). A similar narrative is true for the propensity score match, although the image is not as clear because there is a large amount of prognostic variation between matched

pairs. Thus the unmeasured confounder, *U*, induces systematic differences between paired individuals, even after matching. In simulations a stronger confounder, this visual pattern is exaggerated, whereas it is milder when the confounding is more mild (Supplementary Figure S4). These graphics illustrate visually why a variety of matching approaches still produce biased results when unmeasured confounding is at play, and may be a useful visual complement to sensitivity analyses (Rosenbaum 2005 b).

3.4 Illustration: Randomization-assignment-control plots and instrumental variables

Propensity and prognostic scores are by no means the only characterizations of baseline variation important to a causal question. Extensions of assignment-control plots might visualize some combination of propensity and prognostic score with other axes of variation important to a study design. Here, we consider one candidate: an additional axis summarizing instrumental variation. Briefly, an instrumental variable (IV) is a measured covariate that is associated with the treatment, but that has no effect on the outcome except through the treatment (for an introduction, see (Baiocchi et al. 2014)). IV study designs rest on their own set of assumptions that require care and skepticism (for example, that a valid instrument can be isolated from measured baseline variation). However, unlike propensity score approaches, IV study designs do not require the absence of unmeasured confounding, creating an interesting orthogonality between these methods. Here, we consider an illustration which breaks down the treatment assignment mechanism into propensity score and IV components in order to create the randomization-assignment-control plot, interfacing with methodological discussions of propensity scores and instruments (Bhattacharya and Vogt 2007, Wooldridge 2009, Myers et al. 2011).

Implicit in both instrumental variable and propensity score designs is the supposition that there are two components influencing each individual's treatment assignment: confounding variation and randomizing variation. A subject's "decision" to be treated (or not) is directed by influences that are associated with their likely outcome (which we treat as "confounders") and influences that are unrelated to their outcome (which we may treat as " randomizers"). These components can be further broken down into measured and unmeasured variation. Informally,

$$\textit{treatment assignment}\text{`` = ''}\underbrace{\text{measured confounding variation}}_{\text{propensityscore}}\text{`` + ''}\textit{unmeasured confounding variation}\text{`` + ''}\underbrace{\text{measured randomizing variation}}_{\text{IV}}\text{`` + ''}\textit{unmeasured randomizing variation}$$

Randomizing variation is essential and desirable; it underlies the warrant for inference from randomized trials and well-designed observational studies. Confounding variation – which is inherent to observational studies – generally causes bias and must be addressed by any observational study design. The propensity score methods seek to neutralize measured confounding variation so that all that remains is the implicit randomizing variation (unmeasured confounding variation is assumed to be absent). IV designs seek to directly isolate the measured randomizing variation, so that confounding variation (measured and unmeasured) is balanced by the law of large numbers. Much of the propensity score literature implicitly treats randomizing variation as unmeasured, whereas an explicit

requirement of instrumental variable studies is that some randomizing variation can be isolated from the measured covariates.

Confounding and randomizing variation play distinct roles in an observational study, which motivates visualizing them separately and treating them differently. For instance, observations from theory and simulation suggest that IVs should not be included in propensity score models – even though they are associated with treatment assignment – since this may actually increase the bias and variance of the causal effect estimates in the absence of strong ignorability (Bhattacharya and Vogt 2007, Wooldridge 2009, Myers et al. 2011). As an illustrative example, consider the simulation set-up with unobserved confounding from section 3.3, except that now a new measured covariate, Z, is present in the treatment assignment function as an instrumental variable (IV):

$$\phi(X_i, Z_i) = c_1 X_{i1} + c_2 Z_i + \eta U - c_0,$$

$$\Psi(X_i) = \rho X_{i1} + \sqrt{(1-\rho^2)} X_{i2} + \eta U\,.$$

where c_2, c_1, c_0, η and ρ are constants, and U is an unmeasured confounder. Notice that $\phi(X_i, Z_i)$, in spite of capturing the true assignment process, is no longer an ideal "propensity score" for the purposes of matching. Instead, the ideal (somewhat modified) score, $\tilde{\phi}$ would summarize just the variation in X_1 and omit any variation in Z. In some ways this is a departure from the conventional description of the propensity score, in that $\tilde{\phi}$ should summarize confounding variation and *exclude* randomizing variation (i.e. Z). Instead, Z(the IV) could be summarized in its own *randomization* axis. Figure 5 visualizes of this *randomization-assignment-control* space, which – for simplicity – visualizes each individual projected down onto each pair of axes ($\tilde{\phi} \times \Psi, \Psi \times IV$ and $\tilde{\phi} \times IV$). Like Figure 4C-D, these plots show the *true* match distances in light of the unmeasured confounder, U.

Figure 5 shows randomization-assignment-control plots for two study designs. The first (Figure 5A) performs Mahalanobis distance matching on just X_1 and X_2. This matching pairs individuals based on measured confounding variation (X_1) and variation important to the potential outcome (X_2), ignoring the measured randomizing variation (the IV, Z). The second design (Figure 5B) performs nearfar matching, a matching design that directly uses the instrumental variable (Baiocchi et al. 2012, 2014).

The leftmost panel of Figure 5A is an assignment-control plot. This plot shows the unmeasured confounder, U, at work: although the measured confounding and prognostic variation has been efficiently matched upon, there is no way to ensure that matched individuals are close in terms of U. This results in the " slanting" pattern previously shown in Figure 4D: since matched individuals differ in U in a way that is beyond our control, one individual (the one with the higher U value) will tend to be diagonal of the other. This slant in the assignment-control space is for the most part unavoidable: when all of the measured confounding and prognostic variation is matched for, the contribution of unobserved confounders must remain.

How does randomizing variation, then, protect against the bias from unobserved confounding? Assignment-control plots make this easier to visualize. Recall that there are two visual components to bias from unobserved confounding: (1) matched individuals are vertically distant from one another in the assignment-control plot (2) the treated individual is systematically more often the upper (or, for the opposite direction of bias, lower) individual of the pair. Randomizing variation – measured and unmeasured – may protect against unobserved confounding by disrupting component (2). In the rightmost panel of Figure 5A, the upper (higher prognostic score) member of each matched pair is the treated individual more often than the untreated individual, a pattern that will cause bias in a treatment effect estimate from these matched pairs. This is due to the pressure imposed by the unobserved confounder, *U*. However, the upper member of the match is *occasionally* the untreated individual rather than the treated one. This is possible because of randomizing variation, occasionally overcoming the pressure exerted by unmeasured confounding variation. In this way, the randomizing variation partially disrupts the tendency towards bias when an unobserved confounder is at work (see Supplementary Figure S5 for a schematic). This is why including an IV in a propensity score – or indeed, any matching scheme which seeks to *minimize* IV distances within a matched set – can be harmful: randomizing variation between individuals can actually be helpful, so removing it by matching or regression adjustment makes for worse estimation (Supplementary Figures S6-S9).

The Mahalanobis distance matching example suggests how randomizing variation can be protective against bias even when it is ignored. Well-used IV designs, however, directly leverage measured randomizing variation to combat the influences of unobserved confounding. Nearfar matching designs (Figure 5B) are an IV approach that explicitly takes advantage of measured instrumental variation by pairing individuals who are “near” in important non-instrumental covariates (Figure 5B, left panel), but “far” in terms of a measured instrument (Baiocchi et al. 2012) (Figure 5B, center and right panels). An important nuance to this study design is that paired individuals must be divergent in their IV but need not have the opposite treatment assignment. Instead of comparing treated individuals to untreated ones, the nearfar design compares those “encouraged” into the treatment group (i.e. by the IV) with those who were not (Baiocchi et al. 2010, 2012). This design directly disrupts the systematic tendency for treated individuals to deviate from matched untreated individuals by changing the criteria for who can be matched with whom in light of a measured IV.

Figure 5 is intended as an illustration to facilitate the preceding methodological discussion. Randomization-assignment-control plots need further development before they should be directly applied in general, particularly in designs which use instrumental variation explicitly. In specific, the process of fitting the propensity score may be nuanced in study designs using the IV, because selections of treatment assignment are “post-randomization” information in that they occur after the IV takes effect. Pilot design approaches may be a potential tool to avoid overfitting the propensity score in this scenario. With additional work, a visualization similar to Figure 5 may be an especially useful companion to nearfar matching designs (Baiocchi et al. 2012). This may also be of use for other IV study designs for assessing the relationships between a candidate IV (or a set of weak IVs) and other covariates. More broadly, it is important to emphasize that assignment-control plots can and

should be extended to include other relevant aspects of baseline variation *beyond* propensity and prognostic scores, since these extensions may yield new methodological insights. For example, a related plot may be useful in the randomized experiment setting, wherein the prognostic score is visualized alongside an axis (or axes) summarizing compliance behavior.

3.5 In application: Demonstration and practical considerations

Here, we walk through an illustrative example of assignment-control plots applied to a hypothetical study of cardiothoracic surgery outcomes, and discuss some considerations for using assignment-control plots in practice.

3.5.1 A case study of cardiothoracic surgery outcomes—An ongoing concern in cardiology is that female patients tend to have worse outcomes for various cardiovascular events and surgeries. Recent work suggests that, in settings, patient-physician gender concordance may play a role (Greenwood et al. 2018, Wallis et al. 2022). Greenwood et al. (2018) found that female patients exhibit increased mortality from acute myocardial infarction when their physician is male. Observing that women also exhibit higher operative mortality in coronary artery bypass grafting (CABG) surgeries (Blankstein et al. 2005, Mannacio and Mannacio 2018), we consider the question whether patient-physician concordance plays a role in 30-day mortality for CABG procedures. In this illustrative example, we consider the comparison of 30 day mortality between female patients whose primary surgeons identified as female and those whose primary surgeons did not identify as female in a large data set of coronary artery bypass grafting (CABG) surgeries on Medicare patients from 1998 to 2016. Note that a more thorough study of this question would need to address several alternative hypotheses, for example the possibility that differences in outcomes may be explained by different performances between male and female surgeons (Tsugawa et al. 2017). Some of these nuances are discussed by Greenwood et al. (2018).

Prior to most study design or analysis, diagnostic plots can reveal underlying characteristics of the data (Supplementary Figures S2-S3) that inform next steps. These diagnostics may reveal problematic patterns in the data that the researcher might not have otherwise anticipated. For example, rather than focusing solely on female patients, a researcher might have instead begun this work by pooling surgeries on male *and* female patients, intending to study patient-physician concordance across sexes. However, the diagnostic assignment-control plots and propensity histograms for the data under such an approach are show strong sub-grouping by patient sex (Supplementary Figure 2), suggesting that such an approach might inappropriately pool across two groups (male and female patients) with naturally very different underlying probabilities of assignment to a concordant surgeon (owing to the fact that a large majority of cardiovascular surgeons in the data are men). An assignment-control plot can also characterize outliers; While the propensity score histogram for this data set shows that there is a small group of observations with very high propensity score (Supplementary Figure S3C), the AC plot shows that these outliers are similar in prognostic score to the rest of the observations in the data set, an insight which can inform decisions to trim these observations (Supplementary Figure S3B, also see Zhu et al. (2021) for suggestions). Finally, estimated prognostic scores and propensity scores are mildly correlated (Pearson's correlation between estimated scores ≈ 0.2) suggesting that

patients with higher probability of being assigned a female surgeon based on their baseline characteristics (higher propensity score) also appear to be at higher risk based on those characteristics (higher prognostic score). This may be driven by the fact that operations performed by female surgeons in the data are more likely to involve an older patient or an emergency admission (Supplementary Figure S3A). An applied researcher can use this AC plot to open discussions with colleagues of how *measured* confounding in the data overall may tend to bias towards *underestimation* of the performance of female surgeons unless properly addressed. The stronger the association between propensity and prognostic scores, the more severe this issue may be.

Figure 6 suggests diagnostics for comparing matching schemes on a subset of the data. A Love plot (Figure 6AB) compares the standardized mean differences in each covariate between the treatment groups before and after matching. Assignment-control plots (Figure 6CD) depict how close the matched pairs are in terms of propensity scores and prognostic scores. Comparing the left and right sides of figure 6, we see that adding a propensity score caliper helps ensure that matched pairs are closer in terms of treatment probability based on the measured covariates, but covariate balance overall may be slightly worse compared to Mahalanobis distance matching without a caliper. This is an important trade-off: matches which are close in propensity score allow for identification of the treatment effect after Rosenbaum and Rubin (1983), however overall covariate balance – to one view – is the primary goal of matching methods (see, for example King and Nielsen (2019), Imai and Ratkovic (2014)). While these debates exist in the methodological literature, these visuals facilitate practical discussions about design trade-offs of different matching schemes with collaborators who have varying technical backgrounds.

3.5.2 Considerations for assignment-control plots in practice—Love plots and assignment-control plots are a valuable diagnostic pairing for observational studies. Love plots allow the researcher to consider each covariate separately to identify individual covariate imbalances. However, they do not directly convey information about which covariates are most important to the potential outcomes, which may make it difficult to quickly assess which covariate imbalances are worth extra attention, especially when many covariates are at play. Assignment-control plots allow the researcher to assess the relationship between propensity and prognosis, identify potential violations of assumptions, and assess match quality in terms of propensity and prognosis simultaneously (Table 1).

Importantly, assignment-control plots in application are vulnerable to many of the same weaknesses as propensity and prognostic score methods in general. An assignment-control plot is only as reliable as the propensity and prognostic scores from which it is constructed. If model fit to the analysis set is poor, or if there is unobserved confounding at play (section 3.3), assignment-control plots may be misleading. Future work might consider how uncertainty bounds around propensity and prognostic scores might be estimated, displayed, and interpreted, or how assignment-control plots might be used as a diagnostic for the fitted score models themselves. Additionally, when using the prognostic score, it may be advisable to hold aside a subsample of the untreated observations for estimating the prognostic model in order to avoid overfitting. Aikens et al (Aikens et al. 2020) discuss some considerations

for when a “Pilot Design” such as this is most useful, and when it may be too costly of a data sacrifice.

If the study plan already includes using a prognostic score, there is low cost and potential benefit to generating assignment-control plots. First, no additional data sacrifice is necessary to make the plots if the fitting the prognostic score was already in the study plan. Second, a researcher who has selected a study design that uses prognostic and propensity scores is already making the implicit assumption that the estimated scores are adequate, and checking an assignment-control plot can be a responsible way ensure that nothing anomalous or concerning appears in the marginal or joint distribution of the scores. Assignment-control plots may *also* be useful in scenarios where propensity or prognostic scores are not a planned part of study design, although this requires more judgement regarding whether the data sacrifice to build a prognostic score solely for visualization is a worthwhile trade-off for potential insights gained from the plot.

4 DISCUSSION

Assignment-control plots are nested in a broader conversation about the differing types of baseline variation and their differing significances to a causal question Hansen and Klopfer (2006), King and Nielsen (2019), Bhattacharya and Vogt (2007), Wooldridge (2009), Bang and Robins (2005). The ways these sources of variation interact and how they can be leveraged or mitigated against comprise an implicit focal point of causal inference research, both methodological and applied. The greater clarity our community has in defining and characterizing these concepts, the more effectively we can communicate, teach, design studies, and generate new methodological insights.

A modern shift towards an emphasis on large, passively collected data sets presents a host of challenges and opportunities for researchers interested in causality. As we collect wider data sets with more measured covariates, it will be increasingly important to prioritize the baseline variation that is most important to the causal question – correctly leveraging measured covariates that are useful while deprioritizing measured covariates that are uninformative. The complementary tools of the propensity and prognostic scores are a useful starting place because they summarize two important aspects of baseline variation in the measured covariates: variation associated with the assignment mechanism, and variation associated with the potential outcomes. Sections 3.1, 3.2, and 3.5 illustrate some of the potential uses of assignment-control plots in applied studies. In particular, visual tools such as these may be especially useful in facilitating discussions of study design trade-offs between collaborators from different technical backgrounds.

However, propensity and prognostic scores are not the only important sources of variation in a causal inference study. Other study designs may depend on an instrumental variable, a score summarizing the probability of inclusion in the sample, a summary of baseline variation associated with treatment effect heterogeneity, or compliance information in a randomized encouragement design. Considering how these different types of variation interplay may reveal new methodological possibilities. In section 3.4, we illustrate one extension of the assignment-control plot which adds a ‘randomization’ axis, inviting

a methodological discussion of the role of randomizing variation in protecting against unmeasured confounding.

Assignment-control plots and variations thereof can be thought of as dimensionality reduction tools in that they digest a possibly very large covariate space into a meaningful reduced space that is easier to use and understand. While the possible variations on this theme are numerous, they are fundamentally driven by the same insight: a principled understanding of the different types of baseline variation and their differing significances to a causal question can enable researchers to improve the design of causal inference studies and clarify the way we communicate about them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health under grant T32 LM012409; and by Stanford University under a Stanford Graduate Fellowship in Science and Engineering. The authors would like to thank Dr. Jonathan H. Chen and Dr. Guillaume Basse for their mentorship and helpful thoughts on this work.

References

Abadie A and Imbens GW (2006), 'Large sample properties of matching estimators for average treatment effects', econometrica 74(1), 235–267.

Aikens RC, Greaves D and Baiocchi M (2020), 'A pilot design for observational studies: Using abundant data thoughtfully', Statistics in Medicine.

Aikens RC, Rigdon J, Lee J, Baiocchi M, Goldstone AB, Chiu P, Woo YJ and Chen JH (2021), 'stratamatch: Prognostic Score Stratification using a Pilot Design', The R Journal. Accepted, may change after copy-editing.

Antonelli J, Cefalu M, Palmer N and Agniel D (2018), 'Doubly robust matching estimators for high dimensional confounding adjustment', Biometrics 74(4), 1171–1179. [PubMed: 29750844]

Baiocchi M, Cheng J and Small DS (2014), 'Instrumental variable methods for causal inference', Statistics in medicine 33(13), 2297–2340. [PubMed: 24599889]

Baiocchi M, Small DS, Lorch S and Rosenbaum PR (2010), 'Building a stronger instrument in an observational study of perinatal care for premature infants', Journal of the American Statistical Association 105(492), 1285–1296.

Baiocchi M, Small DS, Yang L, Polsky D and Groeneveld PW (2012), ' Near/far matching: a study design approach to instrumental variables', Health Services and Outcomes Research Methodology 12(4), 237–253. [PubMed: 27087781]

Bang H and Robins JM (2005), 'Doubly robust estimation in missing data and causal inference models', Biometrics 61(4), 962–973. [PubMed: 16401269]

Bertsekas DP and Tseng P (1988), 'Relaxation methods for minimum cost ordinary and generalized network flow problems', Operations Research 36(1), 93–114.

Bhattacharya J and Vogt WB (2007), Do instrumental variables belong in propensity scores?, Technical report, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138.

Blankstein R, Ward RP, Arnsdorf M, Jones B, Lou Y-B and Pine M (2005), 'Female gender is an independent predictor of operative mortality after coronary artery bypass graft surgery: contemporary analysis of 31 midwestern hospitals', Circulation 112(9_supplement), I–323.

Friedman J, Hastie T and Tibshirani R (2010), 'Regularization paths for generalized linear models via coordinate descent', Journal of Statistical Software 33(1), 1–22. [PubMed: 20808728]

Greenwood BN, Carnahan S and Huang L (2018), 'Patient–physician gender concordance and increased mortality among female heart attack patients', Proceedings of the National Academy of Sciences 115(34), 8569–8574.

Greifer N (2021), cobalt: Covariate Balance Tables and Plots. R package version 4.3.1. **URL:** https://CRAN.R-project.org/package=cobalt**URL:**

Hansen BB (2004), 'Full matching in an observational study of coaching for the sat', Journal of the American Statistical Association 99(467), 609–618.

Hansen BB (2008), 'The prognostic analogue of the propensity score', Biometrika 95(2), 481–488.

Hansen BB and Klopfer SO (2006), 'Optimal full matching and related designs via network flows', Journal of Computational and Graphical Statistics 15(3), 609–627.

Hansen LS, Hjortdal VE, Andreasen JJ, Mortensen PE and Jakobsen C-J (2015), '30-day mortality after coronary artery bypass grafting and valve surgery has greatly improved over the last decade, but the 1-year mortality remains constant', Annals of Cardiac Anaesthesia 18(2), 138. [PubMed: 25849679]

Imai K and Ratkovic M (2014), 'Covariate balancing propensity score', Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1), 243–263.

Imbens GW and Rubin DB (2015), Causal Inference in Statistics, Social, and Biomedical Sciences, Cambridge University Press.

King G and Nielsen R (2019), 'Why propensity scores should not be used for matching', Political Analysis 27(4), 435–454.

Leacy FP and Stuart EA (2014), 'On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study', Statistics in medicine 33(20), 3488–3508. [PubMed: 24151187]

Mannacio VA and Mannacio L (2018), 'Sex and mortality associated with coronary artery bypass graft', Journal of thoracic disease 10(Suppl 18), S2157. [PubMed: 30123548]

Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM and Glynn RJ (2011), 'Effects of adjusting for instrumental variables on bias and precision of effect estimates', American journal of epidemiology 174(11), 1213–1222. [PubMed: 22025356]

Petersen ML, Porter KE, Gruber S, Wang Y and Van Der Laan MJ (2012), 'Diagnosing and responding to violations in the positivity assumption', Statistical methods in medical research 21(1), 31–54. [PubMed: 21030422]

Rosenbaum PR (2005 a), 'Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies', The American Statistician 59(2), 147–152.

Rosenbaum PR (2005 b), 'Sensitivity analysis in observational studies', Encyclopedia of statistics in behavioral science 4, 1809–1814.

Rosenbaum PR (2019), DOS2: Design of Observational Studies, Companion to the Second Edition. R package version 0.5.2. **URL**: https://CRAN.R-project.org/package=DOS2**URL**

Rosenbaum PR (2020), 'Modern algorithms for matching in observational studies', Annual Review of Statistics and Its Application 7, 143–176.

Rosenbaum PR and Rubin DB (1983), 'The central role of the propensity score in observational studies for causal effects', Biometrika 70(1), 41–55.

Stuart EA (2010), 'Matching methods for causal inference: A review and a look forward', Statistical science: a review journal of the Institute of Mathematical Statistics 25(1), 1. [PubMed: 20871802]

Tipton E (2014), 'How generalizable is your experiment? an index for comparing experimental samples and populations', Journal of Educational and Behavioral Statistics 39(6), 478–501.

Tsugawa Y, Jena AB, Figueroa JF, Orav EJ, Blumenthal DM and Jha AK (2017), 'Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians', JAMA internal medicine 177(2), 206–213. [PubMed: 27992617]

Wallis CJ, Jerath A, Coburn N, Klaassen Z, Luckenbaugh AN, Magee DE, Hird AE, Armstrong K, Ravi B, Esnaola NF et al. (2022), ' Association of surgeon-patient sex concordance with postoperative outcomes', JAMA surgery 157(2), 146–156. [PubMed: 34878511]

Wooldridge J (2009), Should instrumental variables be used as matching variables, Technical report, Citeseer.

Zhu Y, Hubbard RA, Chubak J, Roy J and Mitra N (2021), ‘Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches’, Pharmacoepidemiology and drug safety 30(11), 1471–1485. [PubMed: 34375473]

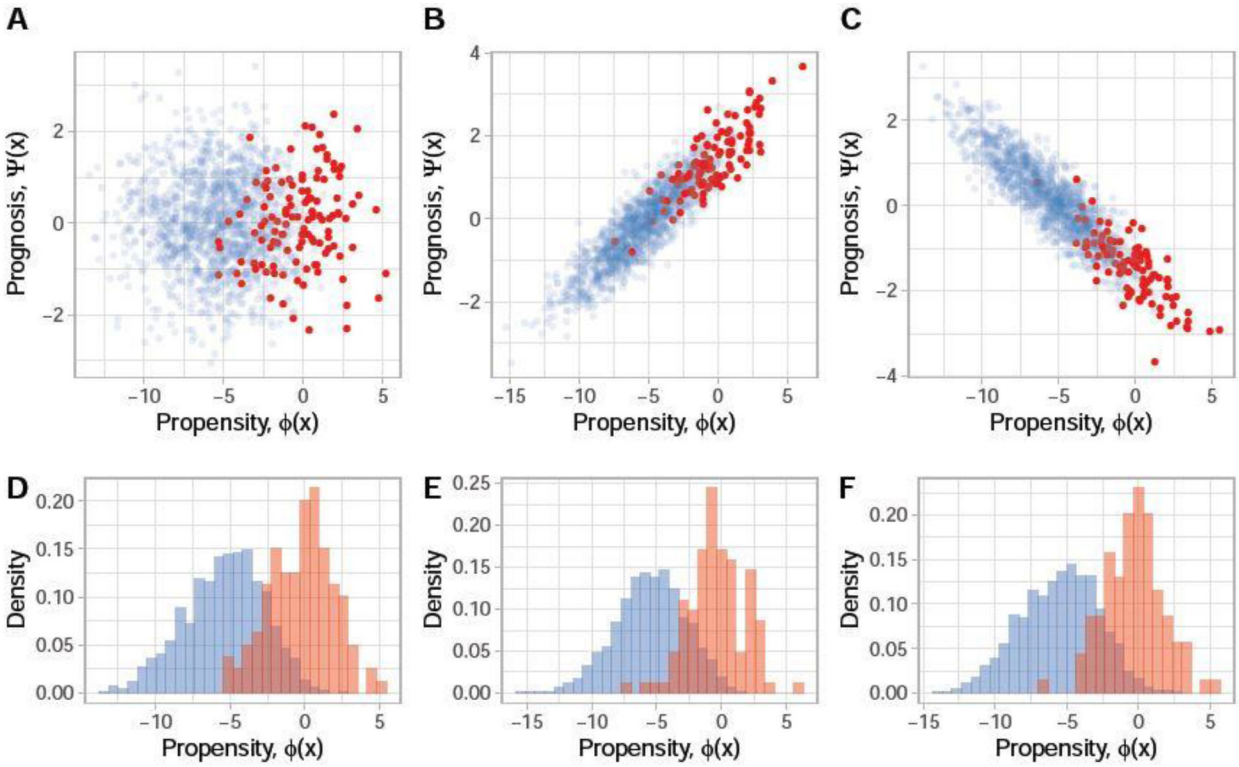

Fig. 1.

Assignment-control plots (A-C) and propensity score density histograms (D-F) for three simulated observational data sets. Red points represent treated observations, blue points represent untreated.

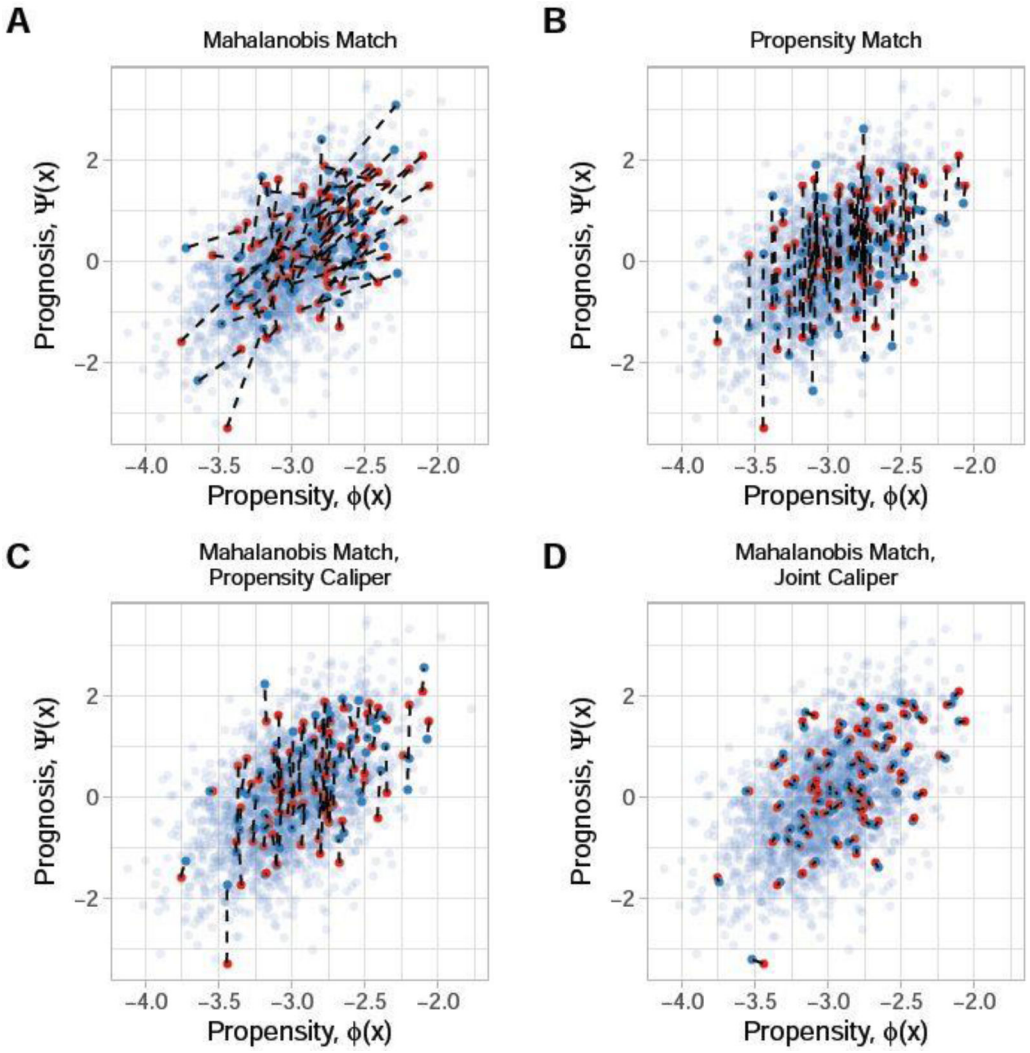

Fig. 2.

Assignment-control plots depicting four different 1-to-1 matching schemes on the same simulated data set. Red points represent treated observations, blue points represent untreated. Dotted lines connect matched individuals. (A) Mahalanobis distance matching, (B), propensity score matching, (C) Mahalanobis distance matching with a propensity score caliper, (D) Mahalanobis distance matching with propensity and prognostic score calipers.

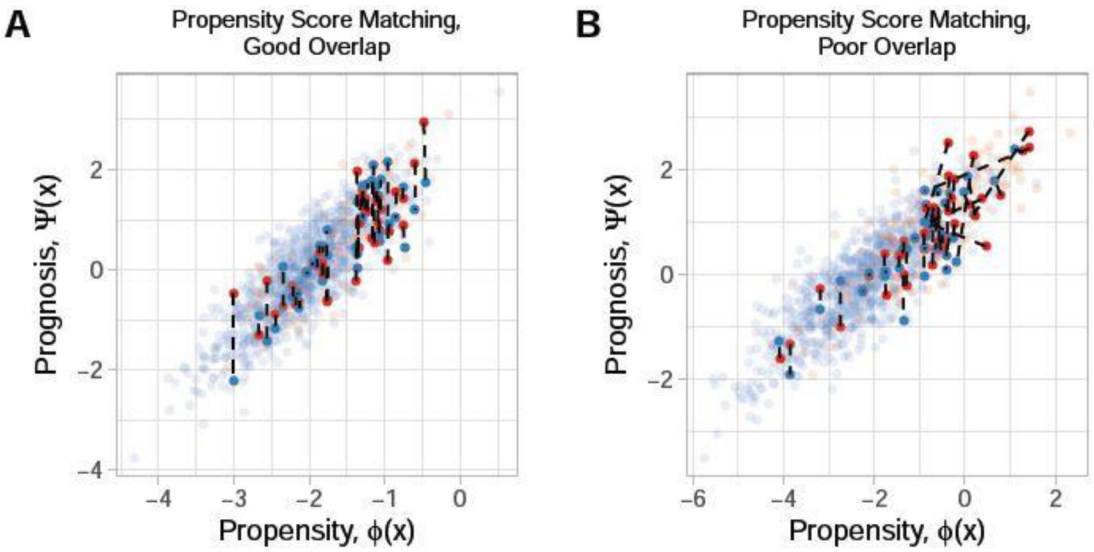

Fig. 3.

Assignment-control plots of propensity score matches in scenarios with good (A) or poor (B) propensity score overlap.

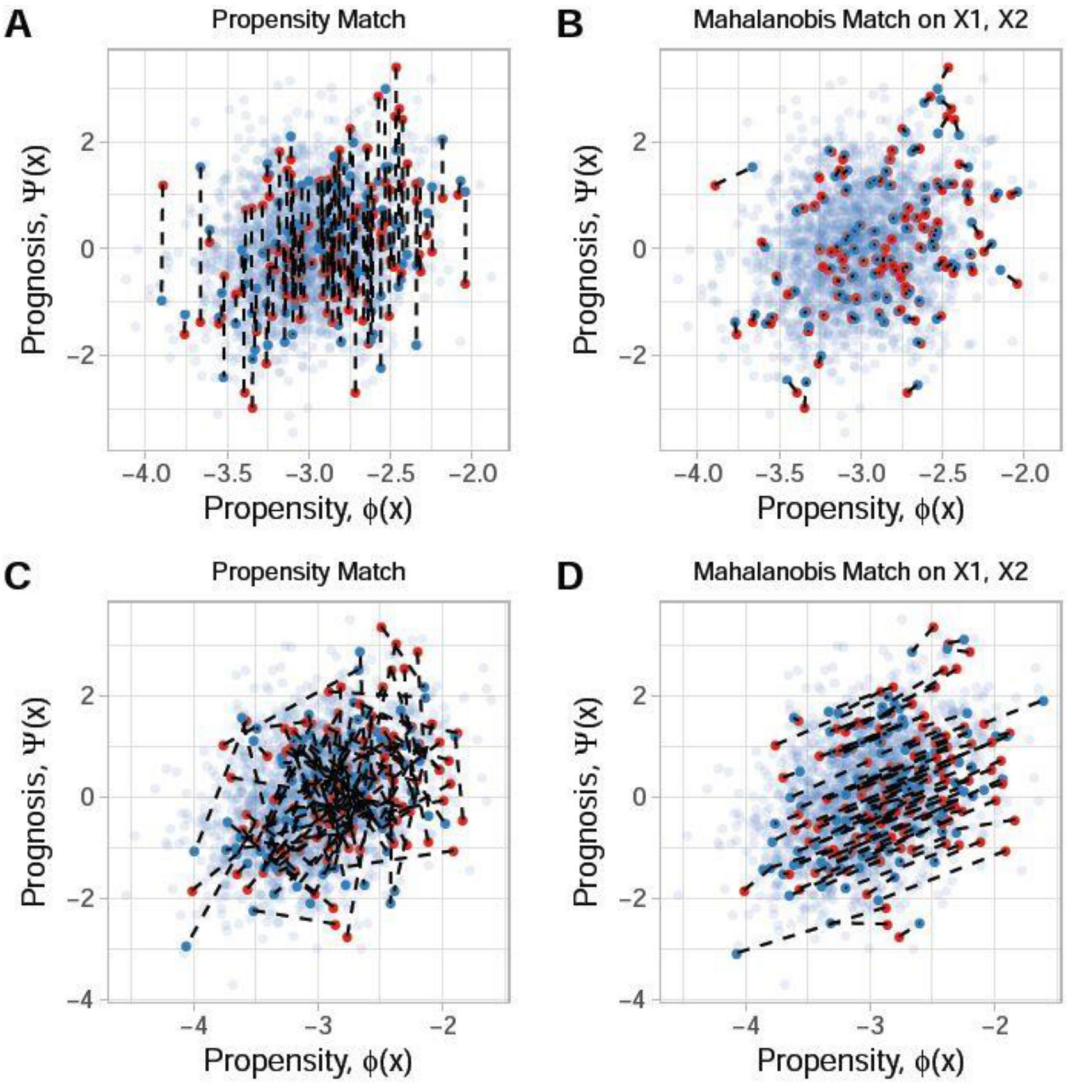

Fig. 4.

Assignment-control plots for two matching schemes on a data set with unobserved confounding. A-B depict the assignment-control space as ascertained without knowledge of the unobserved confounder. C-D depict the true assignment-control space and the true match distances.

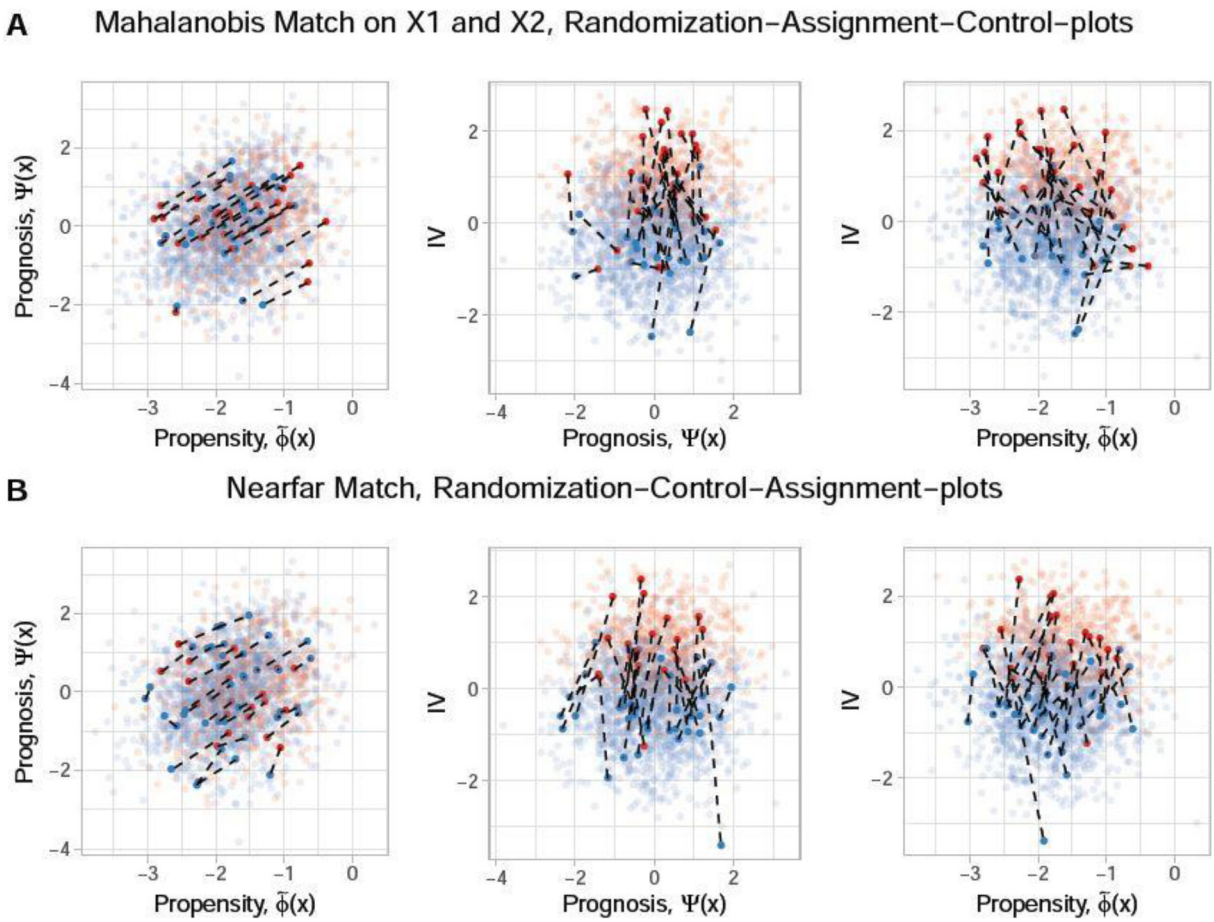

Fig. 5.

Randomization-assignment-control plots. Each panel in a trio depicts a different 2-D projection of the same data set within the randomization-assignment-control space. Red points represent treated observations, blue points represent untreated. Dotted lines connect matched pairs (A) Depicts Mahalanobis distance matching on X_1 and X_2, while (B) depicts a nearfar matching of the same dataset. For visual clarity, only a subsample of 40 matches is shown.

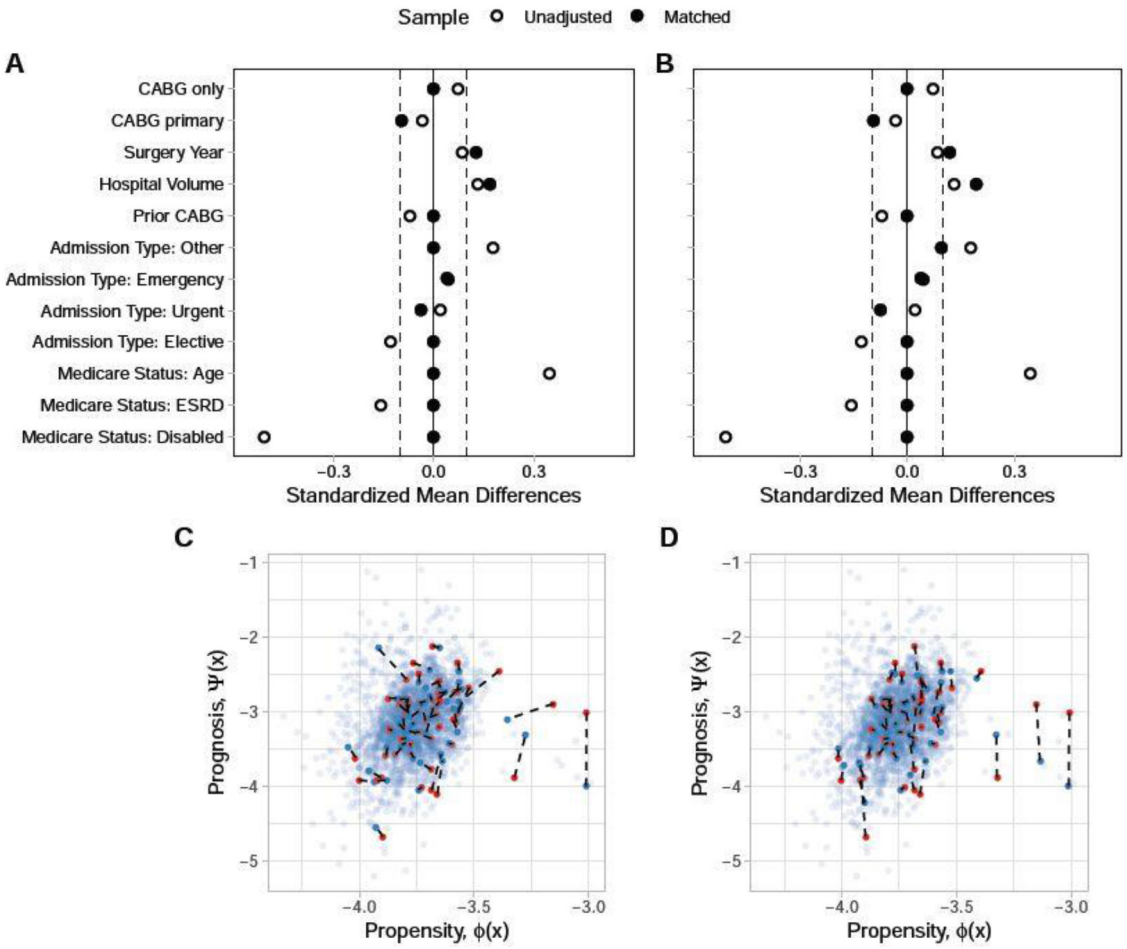

Fig. 6.

Diagnostic plots for a subsample of 2,000 CABG surgeries with and without female primary surgeons. Love plots for a subset of covariates after matching using Mahalanobis distance only (A) versus Mahalanobis distance with a propensity score caliper (B). Assignment-control plots showing matched (dotted lines) pairs of patients with (red) and without (blue) female primary surgeons for the matching scheme without (C) and with the caliper (D). (“ESRD”: “End Stage Renal Disease”)

Table 1

Example characteristics of data visualized in an AC plot and their ramifications for observational study design.

Characteristic	Interpretation	See Also
Weak association between propensity and prognosis	Propensity score matches may be distant on prognostic score. Consider incorporating a prognostic score into matching (e.g. calipers, joint matching).	Figure 1A & 3 Sections 3.1 & 3.2
Positive correlation between propensity and prognosis	Naive analyses will tend to overestimate treatment effect. Consider generalizability of effect estimate, especially if overlap is poor	Figure 1B Section 3.1
Negative correlation between propensity and prognosis	Naive analyses will tend to underestimate treatment effect. Consider generalizability of effect estimate, especially if overlap is poor	Figure 1C Section 3.1
Propensity and prognosis have a nonlinear relationship	Consider reasons for non-linearity and potential for treatment effect heterogeneity	Figure 2A Section 3.1
Observations are strongly subgrouped	This may arise from a highly weighted categorical or binary characteristic in one or both score models. Consider stratification or exact matching within subgroups.	Figure 2B Section 3.1
Propensity overlap is poor	Employ strategies which mitigate against poor overlap. Consider generalizability of effect estimate, especially if propensity and prognosis are highly correlated.	Figures 1 & 2 Sections 3.1 & 3.2
Matches in certain regions of the AC plot are poor	Consider variable k matching or full matching schemes. Ensure that overlap is properly addressed	Figure 3 Section 3.2
Matches are close in propensity score	Decreased bias due to measured confounding.	Figure 2B Section 3.2
Matches are close in prognostic score	Decreased bias due to measured confounding, increased precision, increased power in sensitivity analyses for unobserved confounding.	Figure 2D Section 3.2