Received: 1 December 2017 | Revised: 6 July 2018 | Accepted: 25 July 2018

DOI: 10.1002/bimj.201700294

RESEARCH PAPER

Biometrical Journal

Covariate selection strategies for causal inference: Classification and comparison

Janine Witte[1,2] | Vanessa Didelez[1,2]

[1]Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

[2]Faculty 03: Mathematics/Computer Science, University of Bremen, Germany

Correspondence
Vanessa Didelez, Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Achterstr. 30, 28359 Bremen, Germany.
Email: didelez@leibniz-bips.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DI 2372/1-1

Abstract

When causal effects are to be estimated from observational data, we have to adjust for confounding. A central aim of covariate selection for causal inference is therefore to determine a set that is sufficient for confounding adjustment, but other aims such as efficiency or robustness can be important as well. In this paper, we review six general approaches to covariate selection that differ in the targeted type of adjustment set. We discuss and illustrate their advantages and disadvantages using causal diagrams. Moreover, the approaches and different ways of implementing them are compared empirically in an extensive simulation study. We conclude that there are considerable differences between the approaches but none of them is uniformly best, with performance depending on the chosen adjustment method as well as the true confounding structure. Any prior structural knowledge on the causal relations is helpful to choose the most appropriate method.

KEYWORDS

average causal effect, causal diagram, confounder selection, propensity score matching, variable selection

1 | INTRODUCTION

In epidemiology, a common aim is to estimate causal effects from observational data. This typically requires adjustment for confounding to avoid bias. Hence, a central aim of covariate selection for causal inference is to provide a set of covariates sufficient for confounding adjustment. This is in contrast to covariate selection for prediction, where prediction accuracy is of main concern, or for descriptive modeling, which aims at a sparse representation of the association structure (Shmueli, 2010).

In addition to finding a sufficient adjustment set, covariate selection for causal inference can have other aims such as efficiency. In linear regression, for example, the estimates are most precise when adjusted for predictors of the outcome even when these are not necessary for confounding adjustment. For matching, on the other hand, a central requirement is that the adjustment set be small (see de Luna, Waernbaum, & Richardson, 2011, and references therein). A further aim can be robustness against misspecification of the functional form, or more generally non- or semiparametric estimation. Selecting a small set requires fewer functional forms to be determined. Another idea is to combine outcome- and treatment-oriented selection so that a confounder that is missed out in one selection process has a second chance to be selected in the other one (Belloni, Chernozhukov, & Hansen, 2014). This is related to the double-robust property of adjustment methods that combine treatment and outcome modeling (Bang & Robins, 2005). Finally, in situations where the number of covariates is large compared to the number of observations, the mere reduction of the number of covariates may be an aim in itself.

Considering these different situations and the variety of implied adjustment sets, it becomes clear why so many different approaches to covariate selection have been suggested and also why it can be difficult to decide which method is best for a particular problem. In this paper, we offer some orientation by sketching a classification scheme useful for reasoning about and

TABLE 1 Classification of covariate selection strategies for causal inference

	Preadjustment			Wrapper
	Knowledge-based	**Nonparametric**	**Parametric**	
Minimal approach	• DAGitty algorithm (Textor & Liśkiewicz, 2011) • Common cause criterion (cf. Glymour, Weuve, & Chen, 2008)	• Univariate confounder screening (uniTandY) • CovSel (de Luna et al., 2011)		
Outcome approach		• Univariate outcome screening (uniY) • Model-free variable selection (Li, Cook, & Nachtsheim, 2005) • Random forest variable selection, e.g., Genuer, Poggi, and Tuleau-Malot (2010) and Kursa and Rudnicki (2010)	• Outcome-adaptive lasso (Shortreed & Ertefaie, 2017)	• Optimize outcome model, e.g., AIC, BIC, *p*-value method, validation data
Treatment approach		• Univariate treatment screening	• Optimize treatment model, e.g., AIC, BIC, *p*-value method, validation data	
Union set approach	• Disjunctive cause criterion (VanderWeele & Shpitser, 2011)	• Univariate double screening (uniTorY)	• Double selection (Belloni et al., 2014) • Penalized credible regions (Wilson & Reich, 2014)	
Causal search approach		• EHS algorithm (Entner et al., 2013)		
Estimation approach				• Change in estimate (CIE) • Change in MSE (CI-MSE; Greenland, Daniel, & Pearce, 2016) • Focused confounder selection (FCS; Vansteelandt, Bekaert, & Claeskens, 2012)

Note. Methods in the left-most column assume prior knowledge, the EHS algorithm assumes that all measured covariates are pretreatment, all remaining methods assume that the full set of measured covariates is sufficient.

comparing the different properties of covariate selection strategies when used for causal inference. We begin in Section 2 with key assumptions. Importantly, all selection methods rely on assumptions that can only be justified with subject-matter knowledge. Section 3 outlines our classification: we describe six types of target sets and two selection mechanisms, see also Table 1. In the spirit of Boulesteix, Binder, Abrahamowicz, and Sauerbrei (2018), we conduct an extensive simulation study in Section 4 to compare the principles and methods of selection, carefully separating general approaches from specific implementations by considering the following aspects in turn: First, we investigate how each of the target adjustment sets (minimal, outcome-oriented, etc.) performs in combination with typical adjustment methods (regression, matching, etc.). As the target adjustment sets are only known exactly when (most of) the causal structure is known a priori, we evaluate in a second step the performance of standard implementations (univariate selection, change in estimate [CIE], etc.), for different sample sizes, with regard to their precision and ability to select the desired adjustment set. The results, presented in Section 5, illustrate quantitatively the strengths and weaknesses expected based on theoretical properties; in particular they reveal that there is not a uniformly best selection method. Rather, the target set should be determined based on the method used for adjustment. Not surprisingly, the ability of different methods to find the target sets depends, among other things, on the specific causal structure and the sample size. We conclude the paper with a discussion in Section 6.

2 | SUFFICIENT ADJUSTMENT SETS

The underlying principles of most covariate selection approaches are valid for general treatment types, but for simplicity we consider a binary indicator T of treatment ($T = 1$ when treated, $T = 0$ otherwise). Let further Y be the outcome of interest and $\mathbf{X}^*$ the set of measured covariates. We denote by $\mathbf{X}$ a subset of $\mathbf{X}^*$ with realizations $\mathbf{x} \in \mathcal{X}$. Where appropriate, we use subscripts to $\mathbf{X}$ to indicate the selection method by which $\mathbf{X}$ has been selected from $\mathbf{X}^*$, for example, $\mathbf{X}_{\text{EHS}}$ for a set selected by the Entner–Hoyer–Spirtes algorithm described in Section 3.5. We write $\mathcal{P}(\cdot)$ for distribution functions and $P(\cdot)$ for probability and assume that both are defined on one common population of interest throughout the paper.

The central problem of causal inference from observational data is that the distribution of variables we see as passive observers is different from the distribution we would see if we were able to intervene in the treatment. Therefore, we need a terminology that differentiates between observational and interventional regimes. We use in this paper Pearl's do-notation (Pearl, 2009), where distributions are indexed by $do(Z = z)$ to indicate that Z is set to value z by intervention. For example, $\mathcal{P}(Y; do(T = 1))$ describes the distribution of the outcome Y given treatment is enforced for everyone in the population. In contrast, $\mathcal{P}(Y \mid T = 1)$ describes the distribution of Y in the subpopulation that is observed to receive treatment.

Causal treatment effects can be defined as contrasts between (summaries of) $\mathcal{P}(Y; do(T = 1))$ and $\mathcal{P}(Y; do(T = 0))$. A popular estimand is the average causal effect (ACE),

$$\text{ACE} = \text{E}(Y; do(T = 1)) - \text{E}(Y; do(T = 0))$$

(Imbens, 2004; Lunceford & Davidian, 2004; Schafer & Kang, 2008). An alternative estimand for binary Y is the marginal causal odds ratio (MCOR),

$$\text{MCOR} = \frac{P(Y = 1; do(T = 1))/P(Y = 0; do(T = 1))}{P(Y = 1; do(T = 0))/P(Y = 0; do(T = 0))}$$

(Zhang, 2008). Obviously, in observational studies neither $\mathcal{P}(Y; do(T = t))$ nor any summaries thereof are measured. Identification from observational data requires that the effect can be expressed in $do(\cdot)$-free terms. This is possible when a set $\mathbf{X} \subseteq \mathbf{X}^*$ is available satisfying the following three assumptions:

Assumption 1 (**Pretreatment covariates**).

$$\mathcal{P}(\mathbf{X}; do(T = 0)) = \mathcal{P}(\mathbf{X}; do(T = 1)) = \mathcal{P}(\mathbf{X}).$$

Assumption 1 says that the distribution of the covariates $\mathbf{X}$ is not affected by interventions in the treatment. We call this the pretreatment assumption and say that $X \in \mathbf{X}$ is a pretreatment covariate. Assumption 1 automatically holds if X has been measured prior to T, but it suffices to know that X cannot be affected by T.

Assumption 2 (**Conditional exchangeability**).

$$\mathcal{P}(Y \mid \mathbf{X}; do(T = t)) = \mathcal{P}(Y \mid \mathbf{X}, T = t) \quad \text{for } t = 0, 1.$$

Assumption 2 is the key to identifying causal effects from observational data. Intuitively, it guarantees that conditional on covariates $\mathbf{X}$, association is causation. Assumption 2 is also called "no unobserved confounding" (Robins, 1992). Unless additional experimental data are available or the data happen to contain a structure similar to an instrument (de Luna & Johansson, 2014; Entner, Hoyer, & Spirtes, 2013), it is untestable and therefore needs to be justified by subject-matter knowledge.

Assumption 3 (**Positivity**).

$$P(T = t \mid \mathbf{X} = \mathbf{x}) > 0, \quad \text{for } t = 0, 1 \text{ and all } \mathbf{x} \in \mathcal{X}.$$

Assumption 3, positivity, implies that in sufficiently large samples, both treated and untreated individuals are observed for any given value of the covariates.

It is now shown for $T = 1$ how the interventional distribution $\mathcal{P}(Y; do(T = 1))$ can be expressed by observable terms, using Assumptions 1 and 2 in the second equation:

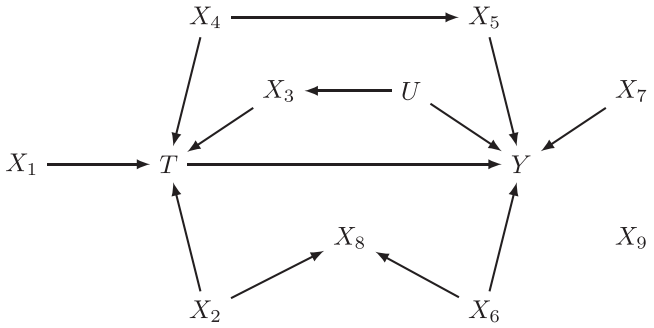

FIGURE 1 Example of causal diagram. T, treatment; Y, outcome; $X_1, \dots, X_9$, observed covariates; U, unobserved covariate. Data generation for setup 1 in Section 4.1 is according to this causal diagram

$$\begin{aligned}
&\mathcal{P}(Y; do(T = 1)) \\
&= \sum_{\mathcal{X}} \mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}; do(T = 1)) P(\mathbf{X} = \mathbf{x}; do(T = 1)) \\
&= \sum_{\mathcal{X}} \mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}, T = 1) P(\mathbf{X} = \mathbf{x}).
\end{aligned}$$

Assumption 3 ensures that $\mathcal{P}(Y \mid \mathbf{X} = \mathbf{x}, T = 1)$ is defined over the whole range of $\mathcal{X}$. The terms in the resulting expression are estimable from observational data. Note that the last equality shows that the intervention distribution can be regarded as a weighted average that illustrates the principle of standardization. This is used in Section 4.3 to estimate the MCOR by standardizing a logistic regression.

We call a set $\mathbf{X} \subseteq \mathbf{X}^*$ satisfying Assumptions 1–3 a *sufficient* adjustment set (Greenland, Pearl, & Robins, 1999). As addressed later, many selection methods assume that $\mathbf{X}^*$ itself is a sufficient adjustment set, and then attempt to reduce this set. We call a sufficient adjustment set *globally minimal* if it has the smallest cardinality among all such sets. We call $\mathbf{X}$ *locally minimal* if it is sufficient and no proper subset of $\mathbf{X}$ is sufficient. Every globally minimal adjustment set is also locally minimal, but not vice versa. Note that none of these sets are necessarily unique.

2.1 | Sufficient adjustment sets and causal diagrams

Causal diagrams are formal but intuitive representations of the underlying causal structure of a problem (Pearl, 2009). Even though it is rare, in practice, that a causal diagram can be fully specified in all detail just based on background knowledge, they are still useful to reason about and illustrate types of adjustment sets. For a short introduction to causal diagrams see Appendix A.1. In the following, we use the terms "ancestor"/"cause" and "parent"/"direct cause" interchangeably. The notion of "direct" cause has to be understood as relative to the variables in the graph, for example, one can often think of unobserved intermediates between variables not shown in the causal diagram. Note that the strong assumptions of a causal diagram can to some extent be relaxed while still allowing identification of sufficient adjustment sets (see Dawid, 2002, for further details).

Assumption 1 (pretreatment covariates) and Assumption 2 (conditional exchangeability) have the following graphical counterparts that can be checked on a given causal diagram (Pearl, 2009):

Assumption 1g (Pretreatment covariates).

Every $X \in \mathbf{X}$ is a nondescendant of T.

Assumption 2g (Backdoor criterion).

All backdoor paths from T to Y are blocked by $\mathbf{X}$.

Thus, $\mathbf{X}$ is a sufficient adjustment set when Assumptions 1–3 or, graphically, Assumptions 1g, 2g, and 3 hold. As an example, consider Figure 1, where $\mathbf{X}^* = \{X_1, \dots, X_9\}$ and U is an unobserved covariate. There are three backdoor paths from T to Y: $T \leftarrow X_4 \rightarrow X_5 \rightarrow Y$, $T \leftarrow X_3 \leftarrow U \rightarrow Y$, and $T \leftarrow X_2 \rightarrow X_8 \leftarrow X_6 \rightarrow Y$. The latter is blocked by the empty set because X_8 is a collider on the path. Examples of sufficient adjustment sets are $\{X_3, X_4\}$, $\{X_1, X_3, X_5, X_9\}$, and $\{X_3, X_5, X_6, X_8\}$. Moreover, the sets $\{X_3, X_4\}$ and $\{X_3, X_5\}$ are both globally and locally minimal.

2.2 | Assumptions underlying covariate selection for causal inference

Assumptions 1–3 define a sufficient adjustment set. We now give assumptions relating such a set and the observed covariates. The majority of selection strategies rely on the following key assumption:

Assumption 4. $\mathbf{X}^*$ itself is a sufficient adjustment set.

In other words, it is assumed that Assumptions 1–3 hold for the full set of measured covariates $\mathbf{X}^*$ (some methods require Assumption 3 only for the actually selected set). Under Assumption 4, selection aims at reducing the number of covariates, for one of the reasons mentioned in Section 1.

A weaker assumption is the following:

Assumption 5.

(i) Every $X \in \mathbf{X}^*$ is a pretreatment covariate.

(ii) $\mathbf{X}^*$ contains a sufficient adjustment set.

Approaches requiring only Assumption 5 but not Assumption 4 are evidently desirable, but we are only aware of one (VanderWeele & Shpitser, 2011). To see the difference between Assumptions 4 and 5, consider the following example: In Figure 1, the set $\{X_1, \dots, X_9\}$ satisfies Assumption 4. However, assume X_2 and X_6 are unobserved, then the set $\{X_1, X_3, X_4, X_5, X_7, X_8, X_9\}$ does not satisfy Assumption 4 (conditioning on X_8 opens a backdoor path) but satisfies Assumption 5 as it contains a sufficient adjustment set.

A further selection strategy, the algorithm proposed by Entner et al. (2013), is sometimes able to infer conditional exchangeability under an even weaker assumption:

Assumption 6.

(i) Every $X \in \mathbf{X}^*$ is a pretreatment covariate.

(ii) Positivity holds for every selected set $\mathbf{X} \subseteq \mathbf{X}^*$.

Note that Assumptions 4–6 are specific to causal inference. When selection is for predictive modeling, confounding is not an issue.

3 | CLASSIFICATION OF COVARIATE SELECTION STRATEGIES

We now describe six different approaches to covariate selection (Sections 3.1–3.6), each corresponding to a different type of target adjustment set. For each approach, there are several proposed methods to implement the approach. They differ mainly in how much prior structural knowledge they assume, and in their validity under different structures. The six approaches correspond to the rows in Table 1. The columns correspond to a second classification criterion, the mechanism of selection, which we describe in Section 3.7.

We illustrate the approaches with the causal diagram of Figure 1, but note that for the majority of selection methods, it is not required that the data-generating mechanism can in fact be represented by a causal diagram. However, even if only incomplete prior knowledge is available, we still recommend to consider a set of plausible diagrams specifically to rule out problematic unobserved quantities, and hence to help justify Assumption 4 or 5. The problem of unobserved covariates is, in fact, a general one but affects the different approaches in different ways; this will be addressed individually below.

3.1 | Minimal approach

3.1.1 | Motivation

Small adjustment sets are advantageous for nonparametric adjustment methods in terms of both bias and variance (de Luna et al., 2011). Especially in the context of matching, small adjustment sets appear desirable as it is then easier to find suitable matches. They are also favorable for regression procedures with continuous covariates because fewer functional forms need to be specified. This first approach can therefore be described as aiming at small, ideally locally or even globally minimal adjustment sets.

3.1.2 | Examples

Given a causal diagram, globally and locally minimal adjustment sets can in principle be read off using the backdoor criterion. For large and complex diagrams, there exist algorithms that list all minimal sets, for example the algorithm given by Textor and Liśkiewicz (2011), implemented in DAGitty (Textor, Hardt, & Knüppel, 2011; *DAGitty algorithm*). In the more realistic situation that the causal structure is not fully known, proposals exist that aim at approximations to such minimal sets. A popular rule requiring partial causal knowledge recommends to adjust for "all common causes of T and Y" (*common cause criterion*; cf. Glymour et al., 2008). Another rule selects all covariates that are associated with the treatment and with the outcome conditional on treatment and do not lie on the causal pathway between treatment and outcome. We call this method *univariate confounder screening (uniTandY)*. It is univariate in the sense that the associations with treatment and outcome are assessed for each covariate separately, not conditionally on other covariates. In contrast, de Luna et al. (2011) suggested to base selection on conditional associations/dependencies and independencies (see also Robins, 1997; VanderWeele & Shpitser, 2011). They describe two algorithms. Roughly speaking, starting from the full set, Algorithm 1 first removes covariates conditionally independent of T given the remaining covariates, then further removes covariates conditionally independent of Y given T and the rest. The alternative Algorithm 2 reverses the role of T and Y. At each stage, the covariates to be removed are chosen so that the number of remaining covariates is as small as possible. The target sets of Algorithms 1 and 2 are each unique but can differ from each other (de Luna et al., 2011). In the example in Figure 1, Algorithm 1 selects the minimal set $\{X_3, X_4\}$ and Algorithm 2 selects the minimal set $\{X_3, X_5\}$. We refer to the general idea as the *CovSel* method, after the associated `R` package. It can be shown that CovSel selects locally minimal adjustment sets under Assumption 4 and additional mild assumptions, given the dependence structure is correctly inferred, which requires a sufficiently large sample size (de Luna et al., 2011).

3.1.3 | Caveats

Selecting minimal sets without knowing the causal diagram is a difficult task in practice. Even the intuitively appealing common cause criterion cannot guarantee under Assumption 4 that a sufficient adjustment set is found, for example, due to unobserved common causes (cf. VanderWeele & Shpitser, 2011). In Figure 1, even though Assumption 4 holds, it will only select $\{X_4\}$, which is insufficient, as U is unobserved. The data-driven methods are of course affected by sampling variation. In particular, univariate confounder screening on the one hand tends to select unnecessarily large sets when the sample size allows many significances. In the example in Figure 1, univariate confounder selection selects X_3, the collider X_8, the unnecessary covariates X_1 and X_2, and both X_4 and X_5 where one of them would suffice. On the other hand, because two tests have to be significant for selection, univariate confounder screening tends to miss important covariates when the sample size is small. Selection of colliders and redundant selection are avoided by CovSel by considering *conditional* (in)dependencies. Except when the true causal diagram is known, none of the aforementioned methods can guarantee globally minimal sets.

3.2 | Outcome approach

3.2.1 | Motivation

The idea of outcome-oriented selection strategies is to determine a sufficient adjustment set that includes strong predictors of the outcome. As a major advantage, adjusting for outcome predictors reduces the standard errors in a variety of settings, including linear outcome regression and propensity score (PS) weighting (Lunceford & Davidian, 2004). Excluding covariates that are conditionally independent of the outcome given treatment and the remaining covariates is a valid strategy that leads to a sufficient adjustment set, provided that Assumption 4 and possibly additional parametric assumptions hold (formally, the conditional independencies relate to potential outcomes as in de Luna et al., 2011, or interventions as in Guo and Dawid, 2010). In terms of a causal diagram, the desired set is sufficient for adjustment and additionally includes all direct causes of the outcome. For the example in Figure 1, the ideal outcome-oriented set would be $\{U, X_5, X_6, X_7\}$. As U is unobserved, the next best set is $\{X_3, X_5, X_6, X_7\}$.

3.2.2 | Examples

Examples are general strategies aiming to select all nonredundant predictors of the outcome. These methods include univariate screening for covariates associated with the outcome either marginally or conditionally on treatment (we refer to the latter as *univariate outcome screening, uniY*). They further include selection methods for parametric regression that aim to *optimize the outcome model* regarding prediction performance or model fit, for example, model selection based on Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), the p-value method (cf. Greenland & Pearce, 2015, for description and criticism), or evaluation of the prediction accuracy using a validation data set. Another

example is *model-free variable selection* as described by Li et al. (2005). Starting from the full set, covariates are removed if conditionally independent of the outcome given treatment and the remaining covariates, based on nonparametric tests. This is the same principle as the first part of CovSel Algorithm 2. Other nonparametric methods include *random forest variable selection* (e.g., Genuer et al., 2010; Kursa & Rudnicki, 2010). An example of how outcome-oriented covariate selection can be combined with treatment modeling is the *outcome-adaptive lasso* where the coefficients of a treatment regression model are penalized inversely proportional to the association of the respective covariates with the outcome in a separate outcome model (Shortreed & Ertefaie, 2017).

3.2.3 | Caveats

Methods focusing on the outcome alone share the drawback that, when used with a small sample size, they might miss covariates that are only weakly associated with the outcome but strongly associated with the treatment and hence still induce confounding bias (Wilson & Reich, 2014). This reflects the fact that covariates important for causal inference are not necessarily as important for outcome prediction, and vice versa. Further, some of the methods have conceptual shortcomings. For example, univariate outcome screening selects redundant covariates and covariates that will not contribute to an improved precision. In the example in Figure 1, the set selected by univariate outcome screening contains X_1 because X_1 is associated with Y conditional on T. However, X_1 is not on a backdoor path and would reduce rather than increase the precision. Multivariate methods avoid this to a certain extent. As an example where even multivariate methods select more covariates than necessary, consider some backdoor path of the form $T \leftarrow X_a \rightarrow X_b \leftarrow U \rightarrow Y$. Although the empty set is sufficient to block this particular path, multivariate methods select $\{X_a, X_b\}$, which is also sufficient but larger than necessary.

3.3 | Treatment approach

3.3.1 | Motivation

As every backdoor path necessarily begins with a direct cause of treatment, the set of all direct causes of T is a sufficient adjustment set (Pearl, 2009), given Assumption 4. Selecting the direct causes of T appears natural when adjustment involves the PS, that is, typically a regression model for treatment given the selected covariates. A main advantage is that selection is clearly separated from any outcome modeling or treatment effect estimation (Rubin, 2001).

3.3.2 | Examples

In principle, all variable selection methods mentioned for the outcome approach can also be used to select predictors of treatment by replacing outcome regression with treatment regression. Basic methods include, for instance, univariate screening for association with the treatment (*univariate treatment screening*) and stepwise regression to *optimize the treatment model* (Weitzen, Lapane, Toledano, Hume, & Mor, 2004).

3.3.3 | Caveats

The treatment model used to estimate the PS should not include causes (or more generally, predictors) of treatment that are not required to block a backdoor path, such as X_1 in Figure 1. It has been shown that adjusting for such unnecessary covariates can lead to a higher variance of causal effect estimates (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Lunceford & Davidian, 2004) and to bias amplification in case of residual unobserved confounding (Bhattacharya & Vogt, 2007; Wooldridge, 2009). Note that treatment-oriented selection is not required to be combined with PS-based adjustment methods but can, for example, be used with regression adjustment. However, this is known to be equally inefficient or biased (Bhattacharya & Vogt, 2007; Myers et al., 2011; Pearl, 2011). Finally, when the sample size is not large enough, association-based methods might miss covariates that are only weakly associated with the treatment but strongly associated with the outcome (Wilson & Reich, 2014).

3.4 | Union set approach

3.4.1 | Motivation

If there is a sufficient adjustment set among the measured pretreatment covariates (Assumption 5), the union of all causes of treatment or outcome is sufficient as well (VanderWeele & Shpitser, 2011). An advantage of this approach is that in the absence of detailed prior knowledge on the causal structure, it is easier for subject-matter experts to justify which covariates are either causes of treatment or outcome. Moreover, when selection is data-driven, the outcome or treatment approaches may miss covariates

that are only weakly associated with the outcome but strongly associated with the treatment, and vice versa. By considering the relationship to the treatment and to the outcome in turn, the selection process gains robustness (Belloni et al., 2014).

3.4.2 | Examples

VanderWeele and Shpitser (2011) suggested to select all causes of treatment or outcome or both (*disjunctive cause criterion*). A data-driven, univariate method pursuing essentially this principle is *univariate double screening* (*uniTorY*), where a covariate is selected if it is associated with treatment or outcome or both (Schafer & Kang, 2008). For high-dimensional problems, Belloni et al. (2014) described *double selection*, where two penalized "nuisance" models are fitted, one for treatment and one for outcome. The union of covariates selected by either penalized model is then used for the causal effect estimation. Another example is the *penalized credible regions* method by Wilson and Reich (2014). Here, Bayesian regression models for treatment and outcome are fitted, and all models within a specified posterior region are defined as "feasible"; within the constrained "feasible" parameter space, the set of covariates with the smallest cardinality is targeted.

3.4.3 | Caveats

The union set approach results in the largest adjustment sets of all approaches considered. This may lead to problems for model fitting and robustness toward misspecification. Replacing the union of causes set by the union set of nonredundant predictors of treatment or outcome, as all methods not based on prior knowledge do, requires again Assumption 4 in order to yield a sufficient adjustment set. Also, the union set might contain strong predictors of treatment that are not needed to avoid confounding bias and therefore might share the disadvantages described for the treatment approach.

3.5 | Causal search approach

3.5.1 | Motivation and example

Entner et al. (2013) described a selection algorithm to which we refer as *EHS algorithm* (for the authors Entner, Hoyer, and Spirtes). It can be viewed as a restricted variant of the Fast Causal Inference (FCI) algorithm for causal search (Spirtes, Glymour, & Scheines, 2000). The EHS algorithm is based on two rules. Rule 1 selects sets $\mathbf{X}_{\text{EHS}} \in \mathbf{X}^*$ so that, for a $X' \in \mathbf{X}^* \setminus \mathbf{X}_{\text{EHS}}$, (i) $X' \not\perp\!\!\!\perp Y \mid \mathbf{X}_{\text{EHS}}$ and (ii) $X' \perp\!\!\!\perp Y \mid \mathbf{X}_{\text{EHS}} \cup T$. It can be shown that sets satisfying Rule 1 are sufficient (Entner et al., 2013). Rule 2 identifies null causal effects and is not discussed here. In the example in Figure 1, assuming that the effect of T on Y is not equal to zero, Rule 1 applies for several combinations of X' and $\mathbf{X}_{\text{EHS}}$, including, for example, $(X' = X_1, \mathbf{X}_{\text{EHS}} = \{X_3, X_5\})$, $(X' = X_2, \mathbf{X}_{\text{EHS}} = \{X_3, X_4, X_5, X_6\})$, and $(X' = X_4, \mathbf{X}_{\text{EHS}} = \{X_2, X_3, X_5, X_7, X_8\})$. The advantage of the EHS algorithm is that in contrast to all aforementioned methods, it is sometimes able to infer that a set of covariates is a sufficient adjustment set based only on Assumption 6, which is considerably weaker than Assumptions 4 and 5.

3.5.2 | Caveats

Although the EHS algorithm returns a list of sufficient adjustment sets in theory, a main disadvantage is that when the dependence structure has to be inferred from data, it likely happens that the rules contradict each other and some amount of user discretion is warranted to interpret the results. Further, when an appropriate X' does not exist the EHS algorithm cannot return any result.

3.6 | Estimation approach

3.6.1 | Motivation

As we are primarily interested in estimating a causal effect, a natural approach is to target sufficient adjustment sets that directly optimize desirable properties of the estimator, especially precision.

3.6.2 | Examples

One can regard selection using the *CIE* criterion as aiming for a low-bias estimator with a minimal covariate set: A benchmark effect is first estimated using all covariates, then covariates are gradually removed until any further removal would result in a change in the estimate of more than, for example, 10%, compared to the benchmark estimate. Greenland et al. (2016) suggested instead to evaluate the *change in the mean-squared error* (*CI-MSE*) of the estimate, which they approximate based on the standard error provided by the regression output. A related and more sophisticated procedure is proposed by Vansteelandt et al. (2012). They described *focused confounder selection* (*FCS*) for logistic regression. Their method takes into account that conditional

effects, such as conditional odds ratios, are not collapsible and therefore focuses on the marginal, regression-standardized (log) odds ratio. The MSE of the marginal effect is estimated using either cross-validation or an asymptotic approximation.

3.6.3 | Caveats

The CIE procedure is a heuristic method that can improve as well as corrupt the properties of the estimator (Greenland & Pearce, 2015). For example, the estimator based on the reduced model can be both more biased and more variable than the benchmark effect. All of the example methods rely on Assumption 4, as they all use the full model to obtain a benchmark estimate. However, even under Assumption 4 there are no guarantees that these methods select sufficient adjustment sets.

3.7 | Mechanism of selection

The six types of target sets described in Sections 3.1–3.6 form our first classification criterion (rows in Table 1). The columns correspond to a second criterion, the mechanism of selection. We distinguish between two general mechanisms, preadjustment and wrapper, and subclassify preadjustment methods into knowledge-based, nonparametric, and parametric methods.

Preadjustment methods separate selection from adjustment. For instance, if selection is completely knowledge-based, it is independent of the adjustment process, which might involve, for example, matching or regression adjustment. The analyst can select different adjustment sets for different adjustment methods (e.g., a minimal set for matching and an outcome-oriented set for regression), but crucially, the selection process is not influenced by the resulting estimate. The same is true for many data-driven selection methods, including all variants of univariate screening. We also classify as preadjustment all methods that perform selection for nuisance models such as the PS. This is in line with the notion that PS estimation is part of the design, not the analysis, of a study (Rubin, 2001).

When using wrapper methods, selection and estimation cannot be separated. Instead, the causal effect is repeatedly estimated with different adjustment sets and one set is selected that optimizes some criterion. The selection procedure is wrapped around the estimation procedure, hence the name (see, e.g., Saeys, Inza, & Larrañaga, 2007, for a similar usage of the term "wrapper"). For example, for CIE the selection criterion is the cardinality of the adjustment set. The CIE method as described in Section 3.6 is a backward procedure, meaning one starts with the full set of covariates, then gradually removes covariates without readding them in later stages. Other types of wrapper algorithms are forward, stepwise, and exhaustive search.

Preadjustment methods have the advantage that they offer a certain amount of safeguard against researchers' discretion because the selection process is not influenced by the estimated causal effect. Further, preadjustment selection is generally flexible, that is, it can be combined with different adjustment methods. In contrast, wrapper methods are specific to one adjustment method, usually a regression model. A disadvantage of all parametric methods, either preadjustment or wrapper, is that one has to assume a functional form for each continuous covariate prior to selection. We also note as an important caveat that whenever data-driven methods are used for selection, inference needs to be adjusted for this selection (Leeb & Pötscher, 2005). For example, resampling the entire selection and estimation process is one way to guarantee valid confidence intervals (Heinze, Wallisch, & Dunkler, 2018). Only few methods are robust toward selection without further provisions, see, for example, Belloni et al. (2014) and Dukes, Avagyan, and Vansteelandt (2018).

4 | SIMULATION SETUP

The aim of our simulation study is twofold: First, we compare the target adjustment sets themselves in the context of typical adjustment methods: (a) outcome regression (combined with standardization for noncollapsible measures, such as odds ratios, when estimating a marginal effect parameter), (b) matching on covariates, (c) matching on the estimated PS, and (d) doubly robust estimation, where both treatment and outcome are modeled and estimators are consistent as long as either model is correctly specified. Doubly robust estimation appears especially attractive when the adjustment set is large, such as with the union set approach, as models are then more likely to be misspecified. In practice, method (a) is often combined with outcome-oriented, method (c) with treatment-oriented, and method (d) with union-set selection. However, it is important to note that all adjustment methods can be combined with any type of adjustment set. Second, we evaluate how well common data-driven methods (see Section 4.2) select their target sets and estimate the target causal effect.

We investigated two general setups, the first following Figure 1 and the second following Figure 2. In setup 1, all causal connections were relatively strong, which made it easy for selection algorithms to detect all relevant associations. In setup 2, in contrast, covariates X_1, X_3, X_5, and X_7 influenced treatment only weakly, and covariates X_2, X_4, X_6, and X_8 influenced outcome only weakly, making it more difficult to detect all associations.

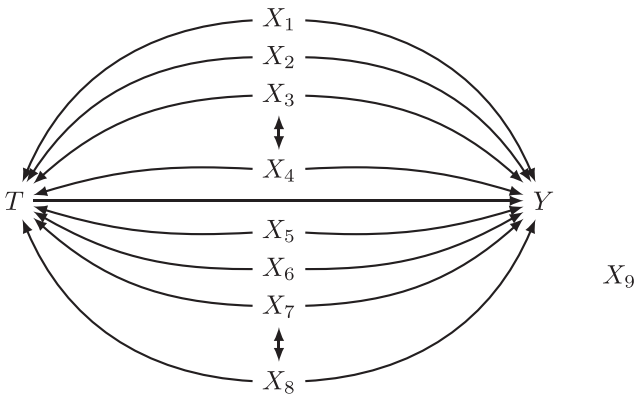

FIGURE 2 Causal diagram used for data generation in setup 2, see Section 4.1. Double-headed arrows indicate correlated error terms, which can be thought of as resulting from an unobserved common cause

For each setup, we investigated two scales of outcome (continuous, binary), three scales of covariates (continuous, mixed scale, binary), two treatment effects (present, absent), and three sample sizes (100, 500, 2,000), resulting in 72 scenarios in total. In the mixed-scale scenarios, the binary covariates were X_3, X_4, X_6, X_7, X_9 in setup 1 and X_1, X_2, X_3, X_4 in setup 2.

In short, the following steps were carried out: (a) Data were generated from the causal diagram in Figure 1 or Figure 2, (b) adjustment sets were determined as the true target sets as well as by applying different selection methods, (c) the causal effect was estimated adjusting, in turn, for the different selected sets. The three steps were repeated 1,000 times. The full code, written for the software package `R` (R Core Team, 2018), is available as Supporting Information on the journal's web page. The online Supporting Information also includes an overview of the `R` packages available for the methods in Table 1.

4.1 | Data generation

The detailed formulas used for data generation are in Appendix A.2. In short, continuous variables were generated according to linear models and binary variables according to logistic models, without interaction effects. Intercepts were chosen such that the prevalence of treatment and outcome, when binary, was about 0.5. In setup 1, for continuous outcome the true ACE is either 0 or 0.5 and an unadjusted analysis has a bias of 0.86/0.58/0.71 for continuous/mixed/binary covariates. For binary outcome, if the treatment is effective the true log(MCOR) is 0.91/0.95/0.88 and an unadjusted analysis has a bias of 0.55/0.37/0.44 for continuous/mixed/binary covariates; if treatment is not effective, the bias is 0.53/0.37/0.47. In setup 2, for continuous outcome the true ACE is either 0 or 0.5 and an unadjusted analysis has a bias of 0.84/1.02/1.15 for continuous/mixed/binary covariates. For binary outcome, the true log(MCOR) is 0.60/0.54/0.51 and an unadjusted analysis has a bias of 0.53/0.61/0.64 for continuous/mixed/binary covariates; if treatment is not effective, the bias is 0.51/0.60/0.63.

4.2 | Implementation of selection methods

The true target adjustment sets for setup 1 are given as follows.

$\mathbf{X}_{\text{min}}$: The minimal target set $\{X_3, X_5\}$ was used.

$\mathbf{X}^*$: The full set $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9\}$ was used.

$\mathbf{X}_Y$: The outcome-oriented target set $\{X_3, X_5, X_6, X_7\}$ was used.

$\mathbf{X}_T$: The treatment-oriented target set $\{X_1, X_2, X_3, X_4\}$ was used.

The following covariate selection strategies were implemented. If not stated otherwise, we used default options for all `R` functions.

AIC: AIC-based selection was implemented as a stepwise (linear or logistic) outcome regression procedure with main effects only, starting with the full set of covariates.

BIC: Analogous to AIC.

Boruta: The `Boruta` package (Kursa & Rudnicki, 2010) was used. All covariates tagged as `Confirmed` by the algorithm were selected into $\mathbf{X}_{\text{Boruta}}$. Boruta is an algorithm for random forest variable selection; covariates are selected based on their so-called variable importance compared to artificially generated noninformative covariates.

CIE: CIE was implemented as a backward (linear or logistic) regression procedure with main effects only. The estimated treatment effect from the full model served as the benchmark estimate. Covariates were then removed one by one such that the CIE, compared to the benchmark estimate, was as small as possible, stopping at a maximum change of 10%. CIE was implemented in two versions: In the coefficient version (CIE_coe), the relevant estimate in each step was the estimated partial regression coefficient of T, corresponding to the effect of T conditional on all other covariates in the model. In the marginal version for binary Y (CIE_mar), regression standardization as described in Section 4.3 was used to estimate the marginal effect of T before assessing the change in the estimate.

CI-MSE: We followed the instructions in Greenland et al. (2016) to implement the CI-MSE method (CI-MSE_coe). In addition, we implemented the variant CI-MSE_mar for binary Y in which the MSE of the marginal treatment effect is estimated in each step, using again standardization, see Section 4.3.

CovSel: For scenarios with only continuous covariates, the package `CovSel` (Häggström, Persson, Waernbaum, & de Luna, 2015) was used. For scenarios with binary covariates, we obtained the source code of the faster `CovSelHigh` (Häggström, 2017; see also Häggström, 2018) from `GitHub`. We modified the code so that continuous covariates were not discretized. We used the modified function with options `method=mmpc, simulate=FALSE, betahat=FALSE`. From the results that were returned by either `cov.sel` or `cov.sel.high`, the sets $\mathbf{X}_{\text{CovSelQ}}$=`Q0` ∪ `Q1`, $\mathbf{X}_{\text{CovSelX.Y}}$=`X.Y0` ∪ `X.Y1`, $\mathbf{X}_{\text{CovSelX.TY}}$=`X.T` ∪ `X.Y0` ∪ `X.Y1`, and $\mathbf{X}_{\text{CovSelZ}}$=`Z0` ∪ `Z1` were extracted, where $\mathbf{X}_{\text{CovSelQ}}$ is the minimal set selected by Algorithm 1, $\mathbf{X}_{\text{CovSelX.Y}}$ is an outcome-oriented set, $\mathbf{X}_{\text{CovSelX.TY}}$ is the union set of the outcome-oriented and a treatment-oriented set, and $\mathbf{X}_{\text{CovSelZ}}$ is the minimal set selected by Algorithm 2.

doubleAIC: AIC selection as described above was performed separately on the outcome model and the treatment model. All covariates selected by either model were selected into $\mathbf{X}_{\text{doubleAIC}}$.

doubleBIC: Analogous to doubleAIC.

FCS: Focused confounder selection (for binary Y) was implemented as a stepwise procedure for standard logistic regression (without PS), starting from the full model. Code for the main selection algorithm was kindly provided by Prof. Vansteelandt.

uniY: For every covariate $X \in \mathbf{X}^*$, Y was regressed on X and T, using linear regression for continuous Y and logistic regression for binary Y. X was selected into the adjustment set $\mathbf{X}_{\text{uniY}}$ if the p-value of X in this model was ≤ 0.05.

uniTorY: For every covariate $X \in \mathbf{X}^*$, T was regressed on X using logistic regression. X was selected into the adjustment set $\mathbf{X}_{\text{uniTorY}}$ if the p-value of X in this model was ≤ 0.05 or $X \in \mathbf{X}_{\text{uniY}}$ or both.

uniTandY: For every covariate $X \in \mathbf{X}_{\text{uniY}}$, T was regressed on X using logistic regression. X was selected if the p-value of X in this model was ≤ 0.05.

4.3 | Estimation of causal effects

Linear regression: The treatment effect was estimated as the partial regression coefficient corresponding to T and the conventional standard error estimate was calculated. If the linear regression model is correct, this corresponds to the ACE.

Logistic regression: Logistic regression was followed by standardization using the package `stdReg` (Sjölander & Dahlqwist, 2017) to obtain an estimate of the log(MCOR): For $t = 0, 1$, the outcome Y was predicted for each individual with its observed covariate values and T set to t. The mean predicted outcomes μ_0 and μ_1 were obtained and the log(MCOR) was estimated as $\log(\mu_1/(1-\mu_1)/\mu_0 * (1-\mu_0))$.

(*PS*) *matching*: Matching was performed using the package `Matching` (Sekhon, 2011) with `estimand=ATE` (corresponding to the ACE). For PS matching, the PS was estimated with a logistic regression model. For continuous Y, the estimated average treatment effect and the Abadie–Imbens standard error were obtained. For binary Y, the matched sample was employed to estimate μ_0 and μ_1 and the marginal log odds ratio was calculated as described above. Note that none of the matching methods were implemented with any pruning of observations.

Doubly robust estimation: For doubly robust estimation, the `iWeigReg` package (Tan & Shu, 2013) was used. This implements a calibrated likelihood estimator that has been shown to outperform other doubly robust estimation procedures (Tan, 2010). The treatment was modeled by logistic regression, the outcome by linear or logistic regression. Both models are correctly specified if the selected covariates contain the parents of treatment and outcome, but they may be misspecified otherwise.

5 | RESULTS

We discuss the results for two of the 72 scenarios: Scenario A is with setup 1, continuous outcome, continuous covariates, effective treatment, and $N = 500$. Scenario B is similar but with binary outcome. The results for the other scenarios are

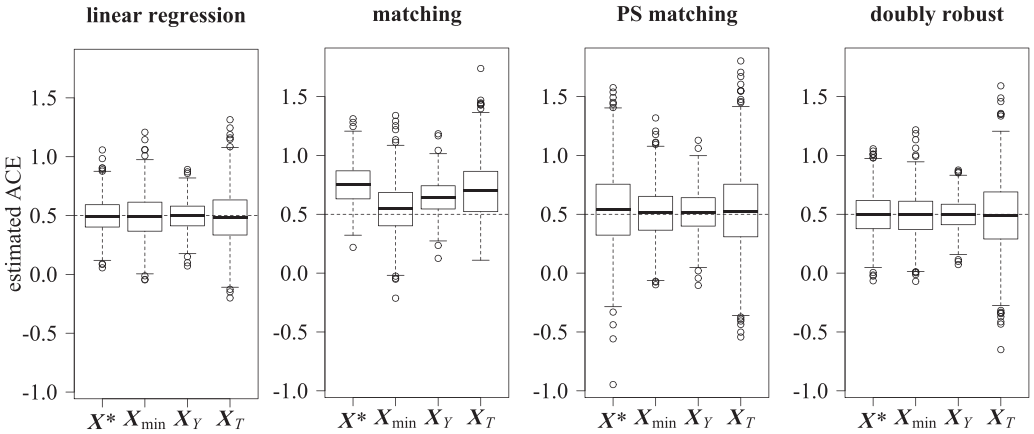

FIGURE 3 Estimated effects in scenario A (continuous outcome, continuous covariates, $N = 500$) adjusted for target sets. Plots show the average causal effect (ACE) estimated by linear regression, matching, propensity score (PS) matching, and doubly robust estimation. The true ACE is 0.5 (dashed line), an unadjusted analysis yields about 1.36

provided as Supporting Information on the journal's webpage. Key findings from the other scenarios are also mentioned in the text.

5.1 | Adjusting for target sets

We start by comparing adjustments with different types of true target adjustment sets. Figure 3 shows box-plots of the estimated ACEs for scenario A with different adjustment methods. Using linear regression, all adjustment sets result in unbiased estimators. The variance is smallest for the outcome-oriented set $\mathbf{X}_Y$ and largest for the treatment-oriented set $\mathbf{X}_T$. In stark contrast, when matching on the covariates, only the minimal set $\mathbf{X}_{\text{min}}$ leads to near unbiased estimators. The bias is largest when adjusting for the full set $\mathbf{X}^*$. The reason is that the more covariates need to be matched on, the worse the matches get with respect to each single covariate. In practice one would prune observations with bad matches; however, as we can see from the improved results, it is not necessary to prune with smaller adjustment sets. Interestingly, the bias for $\mathbf{X}_T$ is notably larger than for $\mathbf{X}_Y$, although these two sets are of the same size. A possible explanation is that covariates strongly associated with treatment tend to be differently distributed in the treatment group versus the control group so that finding good matches is harder. Again, adjusting for $\mathbf{X}_Y$ yields the smallest variance. PS matching results in unbiased or near unbiased estimators for all adjustment sets. The variance is smallest for $\mathbf{X}_Y$, slightly larger for $\mathbf{X}_{\text{min}}$, and considerably larger for $\mathbf{X}^*$ and $\mathbf{X}_T$. The latter confirms previous results (Austin et al., 2007). For doubly robust estimation, similar trends can be observed as for linear regression, with some loss of precision when $\mathbf{X}^*$ or $\mathbf{X}_T$ are used. However, given the greater robustness of this method, it is noteworthy that the loss is only small.

Figure 4 shows the corresponding results of the same analysis for scenario B, that is, binary outcome. For matching and PS matching, the box-plots show very similar trends as in scenario A. When the adjustment method is standardized logistic regression, all estimators are unbiased. The variance is slightly smaller when adjusting for $\mathbf{X}_{\text{min}}$ or $\mathbf{X}_Y$ instead of $\mathbf{X}^*$ or $\mathbf{X}_T$. However, the differences in efficiencies between the types of adjustment sets appear to be quite small.

Varying the sample size or setting the treatment effect to zero does not change the pattern. Further, the trends remain when using all binary covariates. Surprisingly, however, with mixed-scale covariates matching yields close to unbiased estimators for all adjustment sets, see the online Supporting Information.

5.2 | Adjusting for selected sets

Figures 5 and 6 show results obtained by adjusting for sets selected by the different selection methods. The associated tile plots in Figure 7 illustrate how often covariates are selected by which methods. In each plot, the methods are ordered according to their target set. Treatment-oriented selection was not implemented due to inferior results in Section 5. For comparison, we include the box-plot for the full set $\mathbf{X}^*$.

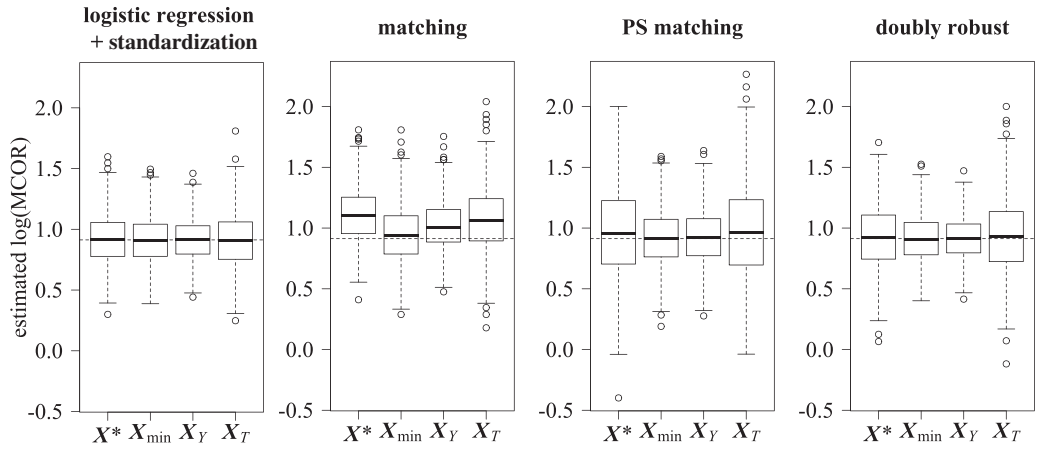

FIGURE 4 Estimated effects in scenario B (binary outcome, continuous covariates, $N = 500$) adjusted for target sets. Plots show the log marginal causal odds ratio (log(MCOR)) estimated by logistic regression with standardization, matching, propensity score (PS) matching, and doubly robust estimation. The true log(MCOR) is about 0.91 (dashed line), an unadjusted analysis yields about 1.46

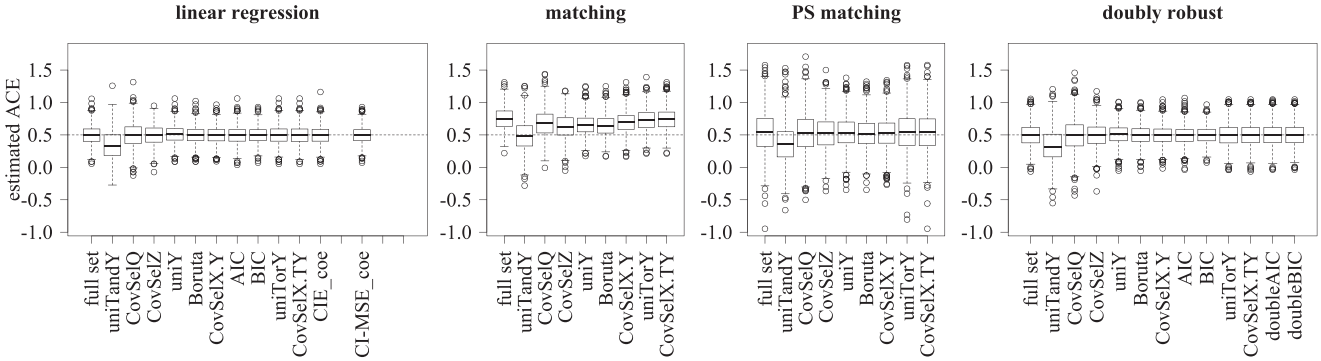

FIGURE 5 Estimated effects in scenario A (continuous outcome, continuous covariates, $N = 500$) adjusted for selected sets. Plots show the average causal effect (ACE) estimated by linear regression, matching, propensity score (PS) matching, and doubly robust estimation. The true ACE is 0.5 (dashed line), an unadjusted analysis yields about 1.36

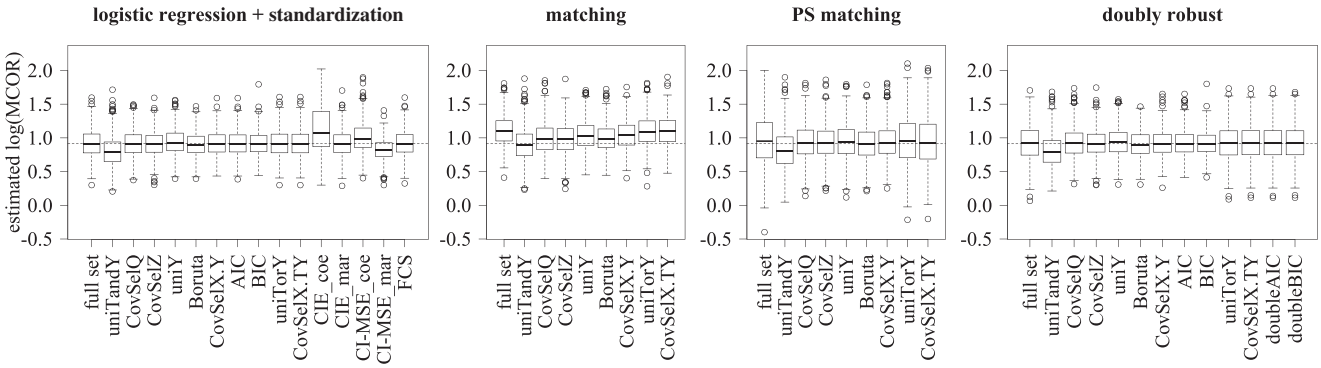

FIGURE 6 Estimated effects in scenario B (binary outcome, continuous covariates, $N = 500$) adjusted for selected sets. Plots show the log marginal causal odds ratio (log(MCOR)) estimated by logistic regression with standardization, matching, propensity score (PS) matching, and doubly robust estimation. The true log(MCOR) is about 0.91 (dashed line), an unadjusted analysis yields about 1.46

5.2.1 | Minimal approach

UniTandY, CovSelQ, and CovSelZ aim at small adjustment sets. For uniTandY, a phenomenon called collider bias can be observed: uniTandY is prone to selecting X_8, a collider on the backdoor path $T \leftarrow X_2 \rightarrow X_8 \leftarrow X_6 \rightarrow Y$. When adjusting for X_8 without also adjusting for either X_2 or X_6, and when all pairwise associations are positive (as in our simulation), a negative association is induced between T and Y, resulting in a negatively biased estimator (Pearl, 2009). This is visible in Figures 5 and

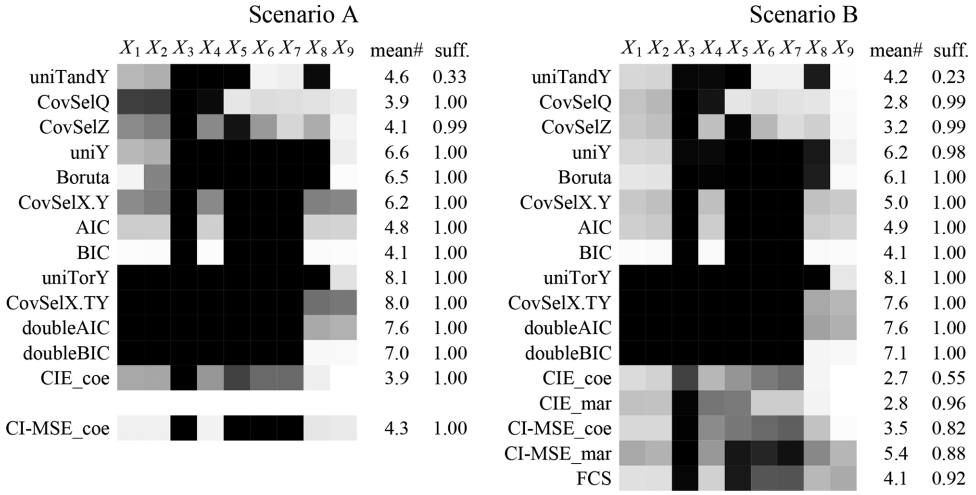

FIGURE 7 Frequencies of covariates selected. The gray level of the tiles is proportional to the number of times the covariate was selected, with white indicating "never selected" and black indicating "always selected." mean#: mean size of selected set; suff.: proportion of times the selected set is sufficient

6 for regression, PS matching, and doubly robust estimation. For matching, the negative bias cancels out with the positive bias from the large number of covariates to be matched on. However, as the size of the set selected by uniTandY strongly depends on the sample size, collider bias cannot be observed for scenario A when the sample size is $N = 100$, as X_8 is then not selected, or when the sample size is $N = 2{,}000$, as X_2 is then selected in addition to X_8 (see online Supporting Information). Ignoring uniTandY, where negative and positive bias cancel out, the smallest bias for matching in scenario A (Figure 5) is achieved by adjusting for the CovSelZ set. The CovSel method performs even better when $N = 2{,}000$ (see online Supporting Information).

5.2.2 | Outcome approach

UniY, Boruta, CovSelX.Y, AIC, and BIC aim at predictors of the outcome. We see in Figure 7 that as expected, uniY tends to select all covariates associated with Y conditional on T. Covariate C_9 is selected in about 5% of cases, reflecting the significance level of the test. The random forest method Boruta selects similar sets but is better able to identify C_9 as an unnecessary covariate. Interestingly, Boruta performs well in combination with matching in many scenarios, including scenario B (Figure 6), although the selected sets are quite large. A possible explanation is that in these scenarios Boruta only rarely selects the strong treatment predictors X_1 and X_2 (Figure 7), for which good matches are especially hard to find. The multivariate methods CovSelX.Y, AIC, and BIC tend to correctly identify $\{X_3, X_5, X_6, X_7\}$ as the direct predictors of Y. The sets selected by AIC and especially BIC are less "noisy" compared to CovSelX.Y, due to the correct parametric assumptions they make. The CovSel method might prove more reliable than AIC and BIC in settings with other than linear influences.

5.2.3 | Union set approach

UniTorY, CovSelX.TY, doubleAIC, and doubleBIC aim at the union set of treatment and outcome predictors. In scenarios A and B, and all other scenarios from setup 1, they tend to select sufficient, but unnecessarily large sets, leading to increased variance and to bias when matching. In setup 2 (see Figure 2), however, where we included many weak associations, the union set methods are the only ones able to find sufficient adjustment sets. As an example, we show in Figure 8 the tile plot for setup 2 with continuous outcome, continuous covariates, effective treatment, and $N = 500$. Only the full set and $\{X_1, \ldots, X_8\}$ are sufficient. The "gaps" in Figure 8 indicate where the selection methods fail to select important covariates due to weak associations.

5.2.4 | Estimation approach

The estimator-oriented selection methods CIE_coe, CIE_mar, CI-MSE_coe, CI-MSE_mar, and FCS are based on parametric regression models (at least in our implementations), hence we show results for their performance only for regression adjustment. In scenario A, selection by CIE_coe (coe for "coefficient version") yields unbiased estimators. The reason is that the full set of covariates is relatively small, so a good estimate is expected without selection (see box-plot "full set" in Figure 5) and the estimate after CIE selection cannot deviate from this by more than 10%, by definition. The same is true for CIE_mar (mar for "marginal version") in scenario B. The coefficient version leads to bias here because the odds ratio is noncollapsible. In general,

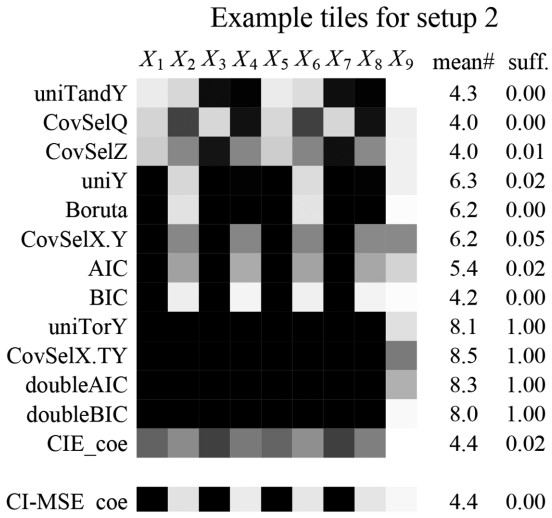

FIGURE 8 Frequencies of covariates selected, for a scenario with setup 2, continuous outcome, continuous covariates, effective treatment, and $N = 500$. The gray level of the tiles is proportional to the number of times the covariate was selected, with white indicating "never selected" and black indicating "always selected." mean#: mean size of selected set; suff.: proportion of times the selected set is sufficient

the CIE method will always perform well when the full model does, especially when the true treatment effect is small. However, this does not mean that the selected covariates are sufficient. For example, in the scenario in Figure 8, CIE selection yields a close to unbiased estimator (see online Supporting Information), yet the selected set is sufficient in only 2% of cases (see Figure 8). Further, it is to be expected that CIE does not perform well when the number of covariates is large so that the estimate from the full model has a large variance. CI-MSE_coe performs well in setup 1 for linear regression, but is unreliable for logistic regression. CI-MSE_mar performs comparable to the union set methods for setup 2 when the sample size is at least 500, but is outperformed in most scenarios by FCS in setup 1 in terms of bias and variance (see online Supporting Information).

5.3 | Coverage

For scenario A, standard ways of estimating the standard error exist; we calculated 95% confidence intervals and assessed how often the true effect of 0.5 was included (Table 2). Note that these standard errors are not corrected for the preceding variable selection. As expected, strong bias leads to severe undercoverage, see, for example, matching on $\mathbf{X}^*$. Moderate undercoverage without bias occurs for doubly robust estimation, especially when the adjustment set is large. For PS matching, the confidence intervals tend to be very wide, resulting in coverage greater than 0.95. For linear regression, mild undercoverage is seen for several adjustment sets, including the target set $\mathbf{X}_Y$. However, we note that 1,000 replications are not sufficient to pin down the mean coverage to the second decimal place, and from the online Supporting Information we find that overall, coverage reaches values close to 0.95 when adjusting for the target sets. When adjusting for the selected sets, the coverage is good in general for $N = 2{,}000$ and decreases with sample size. This is to be expected, as selection is relatively stable when the observations-to-covariates ratio is high (Heinze et al., 2018) so that adjusting for the selected set comes closer to adjusting for a fixed set.

5.4 | Conclusions

In summary, the following conclusions can be drawn from the simulation results. For setup 1 (where only two of nine covariates are necessary for confounding adjustment and the association between causally connected variables is relatively strong), linear regression, logistic regression with standardization, PS matching, and doubly robust estimation worked best when adjustment was for the outcome-oriented target set. Interestingly, however, the results obtained by logistic regression with standardization were quite insensitive to which target set was adjusted for in our simulation. In contrast to all other adjustment methods, matching on the covariates worked best when adjustment was for the minimal set, but was biased when matching was on more than two covariates. In practice, the bias may be reduced by pruning, that is, discarding observations for which no suitable matches can be found. This can be thought of as discarding observations $\mathbf{x}$ for which Assumption 3, positivity, is violated empirically (though not necessarily in the limit). It is important to check for sufficient overlap between the treatment groups regarding the covariates one wants to adjust for, not only for matching but also when using other adjustment methods. In our simulation, regression and PS matching are not affected by insufficient overlap because they are based on correct models that allow for valid extrapolation.

TABLE 2 Coverage analysis for scenario A (continuous outcome, continuous covariates, $N = 500$)

	Linear regression		Matching		PS matching		Doubly robust	
Adjustment set	**Covered**	**Width**	**Covered**	**Width**	**Covered**	**Width**	**Covered**	**Width**
$\mathbf{X}^*$	0.94	0.55	0.68	0.67	0.98	1.48	0.89	0.58
$\mathbf{X}_{\text{min}}$	0.94	0.71	0.94	0.82	0.94	0.87	0.94	0.72
$\mathbf{X}_{Y}$	0.93	0.46	0.82	0.56	0.99	0.88	0.93	0.47
$\mathbf{X}_{T}$	0.95	0.89	0.89	0.99	0.96	1.46	0.90	0.96
$\mathbf{X}_{\text{uniTandY}}$	0.79	0.71	0.92	0.81	0.89	1.11	0.77	0.74
$\mathbf{X}_{\text{CovSelQ}}$	0.96	0.82	0.89	0.92	0.96	1.31	0.92	0.88
$\mathbf{X}_{\text{CovSelZ}}$	0.97	0.70	0.90	0.80	0.97	1.13	0.94	0.73
$\mathbf{X}_{\text{uniY}}$	0.94	0.51	0.83	0.62	0.98	1.17	0.91	0.53
$\mathbf{X}_{\text{Boruta}}$	0.93	0.50	0.84	0.62	0.98	1.13	0.92	0.52
$\mathbf{X}_{\text{CovSelX.Y}}$	0.94	0.50	0.76	0.61	0.99	1.13	0.92	0.52
$\mathbf{X}_{\text{AIC}}$	0.92	0.48	–	–	–	–	0.91	0.49
$\mathbf{X}_{\text{BIC}}$	0.92	0.47	–	–	–	–	0.93	0.47
$\mathbf{X}_{\text{uniTorY}}$	0.94	0.54	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{CovSelX.TY}}$	0.94	0.54	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{doubleAIC}}$	–	–	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{doubleBIC}}$	–	–	–	–	–	–	0.88	0.59
$\mathbf{X}_{\text{CIE\_coe}}$	0.96	0.63	–	–	–	–	–	–
$\mathbf{X}_{\text{CI-MSE\_coe}}$	0.92	0.47	–	–	–	–	–	–

Note. Shown are the proportion of times the true effect of 0.5 was included in the 95% confidence interval (Covered) and the mean width of the interval (Width) when adjusting for the different adjustment sets using linear regression, matching, propensity score (PS) matching, or doubly robust estimation.

We confirmed that univariate confounder selection (uniTandY) is prone to select colliders and that the CIE and the CI-MSE procedure must be used in their marginal versions when the effect of interest is a marginal effect. An important result is that for setup 2 (where eight of nine covariates are necessary for sufficient adjustment but each covariate is responsible for only a small amount of confounding), only methods pursuing the union set approach reliably selected sufficient sets in our simulation. This is unfortunate as the union set as a target set generally leads to inefficient estimation and bias when matching is used. Hence, the question of which approach to use can best be answered based on a priori knowledge of the structure and magnitude of the causal relationships between all variables. If such knowledge is lacking, and if avoiding confounding bias, not efficiency, is the main concern then the union set approach appears to be the safest bet, unless matching is used.

6 | DISCUSSION

In this paper, we distinguished six general approaches to covariate selection for causal inference based on the type of target adjustment set. Common theoretically founded and heuristic methods for implementing these approaches were compared with regard to their theoretical as well as empirical properties. It becomes clear that most selection methods aim at covariate *reduction* rather than selection because they assume that the full set of covariates is a sufficient adjustment set (Assumption 4). Moreover, we argued, and illustrated with simulated data, that different adjustment methods need different types of adjustment sets.

For non- or semiparametric methods, especially matching, small or minimal adjustment sets are clearly desirable. If the underlying causal diagram is known, these can be determined with the DAGitty algorithm. Otherwise, under Assumption 4, CovSel can be recommended while neither the common cause criterion nor univariate confounder screening should be used.

Under Assumption 4, selecting all nonredundant outcome predictors, that is, deselecting variables that are conditionally independent of the outcome given the set of included covariates and treatment increased efficiency not only for regression adjustment but also for PS matching and doubly robust estimation in our simulation. Here many of the data-driven approaches, explicitly or implicitly testing for this type of conditional independence, performed well; univariate outcome screening cannot be recommended.

The treatment-oriented approach cannot be recommended as it is outperformed by all other approaches for all methods of adjustment considered. Note an important difference to the outcome approach: including strong treatment predictors not needed

to adjust for confounding is harmful regarding efficiency of all adjustment methods, and can amplify bias when there is residual unobserved confounding. In contrast, and under Assumption 4, including strong outcome predictors, even if not needed to avoid confounding bias, can still increase efficiency. The main reason to use the treatment approach would be to separate covariate selection completely from modeling of the outcome, for example, to avoid postselection bias.

The disjunctive cause criterion has the advantage that it leads to valid adjustment sets under the weaker Assumption 5 without requiring full knowledge of the underlying causal diagram. However, in many situations, one will not even have the expert knowledge to identify the causes of treatment or outcome. Data-driven methods then require again Assumption 4. The union set approach appears most useful in situations where there are many weak confounders, and where avoidance of confounding bias is the primary concern over efficiency. With the resulting typically large size of adjustment set one may be particularly interested in approaches that are robust toward model misspecification.

In the absence of any prior knowledge to justify Assumptions 4 or 5, the EHS algorithm is an interesting alternative but has not demonstrated its practical use in real-life data examples yet.

There are a number of limitations to our investigation. First, the list of example methods we mention is, of course, not exhaustive. Among others, we did not consider methods that combine two or more approaches, such as the combination of the disjunctive cause criterion with model-free backward selection (VanderWeele & Shpitser, 2011), combinations of causal diagrams and CIE (Evans, Chaix, Lobbedez, Verger, & Flahault, 2012; Weng, Hsueh, Messam, & Hertz-Picciotto, 2009), or the adjustment uncertainty algorithm by Crainiceanu, Dominici, and Parmigiani (2008) combining outcome- and treatment-oriented selection with the CIE criterion. Also, we did not consider model averaging approaches as described in Wang, Parmigiani, and Dominici (2012), Zigler and Dominici (2014), and Talbot, Lefebvre, and Atherton (2015).

In our simulation study, we generated data according to linear and logistic models. In practice, other functional forms for relations between variables may be more plausible. Especially when the covariates are continuous, selection of the covariates itself is only one half of the problem, the other half being model specification, for example, selection of higher order terms and interaction terms. One approach is to consider such terms as additional covariates (Belloni et al., 2014). Another approach selects and adjusts nonparametrically, for example, combining CovSel selection with matching on the covariates.

Another important point we touched only briefly is postselection inference. When effects are estimated from the same data as used to select covariates, the standard errors, confidence intervals, p-values, etc. reported by software are inappropriate. Although in our simulation undercoverage was primarily driven by bias, the effects of postselection inference will be more pronounced with more covariates. For this reason, among others, Heinze et al. (2018) advised to refrain from data-driven selection of covariates altogether when the number of covariates is small to moderate. Although this advice is in principle also sensible, under Assumption 4, when the aim is causal inference, the set of potential confounders will often not be small.

Importantly, all selection methods for causal inference rely on untestable assumptions. The assumption that the full set of measured covariates is sufficient for confounding adjustment, as assumed by the majority of methods, is strong and can only be justified by subject-matter knowledge. Although ideally, one would specify a causal diagram to identify the exact desired target adjustment set (Hernán, Hernández-Díaz, Werler, & Mitchell, 2002), this is often not practical. An understanding of the strengths and weaknesses of alternative selection methods is therefore crucial; our classification and comparison contribute to such an understanding.

ACKNOWLEDGMENTS

We thank two referees and the associate editor for their valuable comments and suggestions. We gratefully acknowledge financial support of the German Research Foundation (DFG – Project DI 2372/1-1).

CONFLICTS OF INTEREST

The authors have declared no conflict of interest.

ORCID


Janine Witte http://orcid.org/0000-0003-0346-2633

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734–753.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–972.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*, 608–650.

Bhattacharya, J., & Vogt, W. B. (2007). *Do instrumental variables belong in propensity scores?* (NBER Technical Working Paper No. 343). Cambridge, MA: National Bureau of Economic Research. Revised 2009. Retrieved from http://www.nber.org/papers/t0343

Boulesteix, A.-L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, *60*, 216–218.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149–1156.

Crainiceanu, C. M., Dominici, F., & Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika*, *95*, 635–651.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, *70*, 161–189.

de Luna, X., & Johansson, P. (2014). Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, *2*, 187–199.

de Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*, 861–875.

Dukes, O., Avagyan, V., & Vansteelandt, S. (2018). High-dimensional doubly robust tests for regression parameters. Preprint arXiv:1805.06714v2.

Entner, D., Hoyer, P. O., & Spirtes, P. (2013). Data-driven covariate selection for nonparametric estimation of causal effects (pp. 256–264). In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), Scottsdale, AZ.

Evans, D., Chaix, B., Lobbedez, T., Verger, C., & Flahault, A. (2012). Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. *BMC Medical Research Methodology*, *12*, 156.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*, 2225–2236.

Glymour, M. M., Weuve, J., & Chen, J. T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. *Neuropsychology Review*, *18*, 194–213.

Greenland, S., Daniel, R., & Pearce, N. (2016). Outcome modelling strategies in epidemiology: Traditional methods and basic alternatives. *International Journal of Epidemiology*, *45*, 565–575.

Greenland, S., & Pearce, N. (2015). Statistical foundations for model-based adjustments. *Annual Review of Public Health*, *36*, 89–108.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*, 37–48.

Guo, H., & Dawid, A. P. (2010). Sufficient covariates and linear propensity analysis (pp. 281–288). Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy.

Häggström, J. (2017). CovSelHigh: Model-free covariate selection in high dimensions [R package version 1.1.1]. Retrieved from https://CRAN.R-project.org/package=CovSel

Häggström, J. (2018). Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, *74*, 389–398.

Häggström, J., Persson, E., Waernbaum, I., & de Luna, X. (2015). CovSel: An R package for covariate selection when estimating average causal effects. *Journal of Statistical Software*, *68*, 1–20.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*, 431–449.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*, 176–184.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*, 1–13.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, *21*, 21–59.

Li, L., Cook, D. R., & Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 285–299.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*, 1213–1222.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*, 1223–1227.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robins, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, *79*, 321–334.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 69–117). New York: Springer.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, *42*, 1–52.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*, 1111–1122.

Sjölander, A., & Dahlqwist, E. (2017). stdReg: Regression standardization. [R package version 2.2.0]. Retrieved from https://CRAN.R-project.org/package=stdReg

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: MIT press.

Talbot, D., Lefebvre, G., & Atherton, J. (2015). The Bayesian causal effect estimation algorithm. *Journal of Causal Inference*, *3*, 207–236.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97*, 661–682.

Tan, Z., & Shu, H. (2013). iWeigReg: Improved methods for causal inference and missing data problems [R package version 1.0]. Retrieved from https://CRAN.R-project.org/package=iWeigReg

Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, *22*, 745.

Textor, J., & Liśkiewicz, M. (2011). Adjustment criteria in causal diagrams: An algorithmic perspective (pp. 681–688). Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), Barcelona, Spain.

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*, 1406–1413.

Vansteelandt, S., Bekaert, M., & Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, *21*, 7–30.

Wang, C., Parmigiani, G., & Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, *68*, 661–671.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, *13*, 841–853.

Weng, H.-Y., Hsueh, Y.-H., Messam, L. L. McV., & Hertz-Picciotto, I. (2009). Methods of covariate selection: Directed acyclic graphs and the change-in-estimate procedure. *American Journal of Epidemiology*, *169*, 1182–1190.

Wilson, A., & Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, *70*, 852–861.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? (Working Paper). East Lansing, MI: Michigan State University. Retrieved from http://econ.msu.edu/faculty/wooldridge/docs/treat1r6.pdf

Zhang, Z. (2008). Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics—Theory and Methods*, *38*, 309–321.

Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, *109*, 95–107.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*. 2018;1–20. https://doi.org/10.1002/bimj.201700294

APPENDIX

A.1 Introduction to causal diagrams

Causal diagrams are statistical models associated with *directed acyclic graphs* (DAGs). In a DAG, variables are represented by *nodes* and causal relationships are represented by arrows, also called *directed edges*. A directed edge from node A to node B, $A \rightarrow B$, means that A has a causal effect on B that is not mediated by any other variable in the DAG. A is then called a *parent* or *direct cause* of B. Although an edge may represent a zero effect, the absence of an edge always indicates that there is no effect.

Any sequence of nodes joined by edges in a DAG is called a *path*, regardless of how the edges are oriented. A path is *directed* if all edges have the same direction. If there is a directed path from A to B, then A is called an *ancestor* or *cause* of B and B is called a *descendant* of A. A directed path from a node to itself is called a *cycle* and is not allowed in a DAG. If A is a nondescendant of B, a backdoor path between A and B is defined as a path from A to B that starts with an arrowhead at A. A *collider* on a path is a node at which two arrowheads collide with respect to that path, $\rightarrow C \leftarrow$.

Causal diagrams have to satisfy the condition of causal sufficiency demanding that every common cause of two variables in the DAG is a node in the DAG as well. A causal diagram not only illustrates causal relations, but also represents the conditional independence structure between all variables in the DAG. A conditional independence between variables A and B given variables $\mathbf{C}$ can be read off the graph by checking whether the nodes in $\mathbf{C}$ *block* every path between nodes A and B. A path between A and B is blocked by $\mathbf{C}$ if (a) it contains a noncollider in $\mathbf{C}$ or (b) it contains a collider such that neither the collider itself nor any descendant thereof are in $\mathbf{C}$. The special case of marginal independence between variables A and B can be seen by checking whether the empty set blocks every path between A and B. This is only possible if there is no path or every path contains a collider. Although blocked paths guarantee conditional independencies, an unblocked path means that a (conditional) association is possible.

A.2 Data-generating mechanisms

In all scenarios, continuous covariates with no incoming arrows were drawn from the standard normal distribution. Binary covariates with no incoming arrows were drawn from the Bernoulli distribution with probability $1/(1+\exp(2))$. In setup 2, the correlated covariates (X_3, X_4) were generated from a bivariate normal distribution such that both had a standard normal distribution marginally and their covariance was 0.5. In the scenarios that specified (X_3, X_4) as binary, they were discretized using 0 as a threshold. The same applies for (X_7, X_8). The remaining covariates were generated according to the following formulas.

In setup 1, for binary X_3, $\mathrm{P}(X_3 = 1) = 1/(1+\exp(2-\beta_{3.U}U))$. For continuous X_3, $X_3 = \beta_{3.U}U + \varepsilon_3$ with $\varepsilon_3 \sim \mathcal{N}(0,1)$, and X_3 was standardized afterward to have zero mean and unit variance. For binary X_5, $\mathrm{P}(X_5 = 1) = 1/(1+\exp(2-\beta_{5.4}X_4))$; for continuous X_5, $X_5 = \beta_{5.4}X_4 + \varepsilon_5$ with $\varepsilon_5 \sim \mathcal{N}(0,1)$ followed by standardization. For binary X_8, $\mathrm{P}(X_8 = 1) = 1/(1+\exp(2-\beta_{8.2}X_2 - \beta_{8.6}X_6))$; for continuous X_8, $X_8 = \beta_{8.2}X_2 + \beta_{8.6}X_6 + \varepsilon_8$ with $\varepsilon_8 \sim \mathcal{N}(0,1)$ followed by standardization. The treatment was generated according to $\mathrm{P}(T = 1) = 1/(1+\exp(\beta_0^T - \beta_{T.1}X_1 - \beta_{T.2}X_2 - \beta_{T.3}X_3 - \beta_{T.4}X_4))$ and the outcome according to $Y = \beta_{Y.U}U + \beta_{Y.5}X_5 + \beta_{Y.6}X_6 + \beta_{Y.7}X_7 + \beta_{Y.T}T + \varepsilon_Y$ with $\varepsilon_Y \sim \mathcal{N}(0,1)$ or $\mathrm{P}(Y = 1) = 1/(1+\exp(\beta_0^Y - \beta_{Y.U}U - \beta_{Y.5}X_5 - \beta_{Y.6}X_6 - \beta_{Y.7}X_7 - \beta_{Y.T}T))$, respectively. The treatment effect $\beta_{Y.T}T$ was 0.5 when treatment had an effect and the outcome was continuous, 1.5 when treatment had an effect and the outcome was binary, and 0 otherwise. The other parameter values varied according to the scale of the covariates: For continuous covariates, $\beta_{3.U} = \beta_{5.4} = \beta_{8.2} = \beta_{8.6} = \beta_{T.1} = \beta_{T.2} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.U} = \beta_{Y.5} = \beta_{Y.6} = \beta_{Y.7} = 1$, $\beta_0^T = 0$, and $\beta_0^Y = 0.5$. For mixed-scale covariates, $\beta_{3.U} = \beta_{8.2} = \beta_{T.1} = \beta_{T.2} = \beta_{Y.U} = \beta_{Y.5} = 1$, $\beta_{5.4} = \beta_{8.6} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.6} = \beta_{Y.7} = 3$, $\beta_0^T = 0.7$, and $\beta_0^Y = 1.1$. For binary covariates, $\beta_{3.U} = \beta_{5.4} = \beta_{8.2} = \beta_{8.6} = \beta_{T.1} = \beta_{T.2} = \beta_{T.3} = \beta_{T.4} = \beta_{Y.U} = \beta_{Y.5} = \beta_{Y.6} = \beta_{Y.7} = 3$, $\beta_0^T = 1.4$, and $\beta_0^Y = 1.9$.

In setup 2, $\mathrm{P}(T = 1) = 1/(1+\exp(\beta_0^T - \beta_{T.1}X_1 - \beta_{T.2}X_2 - \beta_{T.3}X_3 - \beta_{T.4}X_4 - \beta_{T.5}X_5 - \beta_{T.6}X_6 - \beta_{T.7}X_7 - \beta_{T.8}X_8))$ and $Y = \beta_{Y.1}X_1 + \beta_{Y.2}X_2 + \beta_{Y.3}X_3 + \beta_{Y.4}X_4 + \beta_{Y.5}X_5 + \beta_{Y.6}X_6 + \beta_{Y.7}X_7 + \beta_{Y.8}X_8 + \beta_{Y.T}T + \varepsilon_Y$ with $\varepsilon_Y \sim \mathcal{N}(0,1)$ or $\mathrm{P}(Y = 1) = 1/(1+\exp(\beta_0^Y - \beta_{Y.1}X_1 - \beta_{Y.2}X_2 - \beta_{Y.3}X_3 - \beta_{Y.4}X_4 - \beta_{Y.5}X_5 - \beta_{Y.6}X_6 - \beta_{Y.7}X_7 - \beta_{Y.8}X_8 - \beta_{Y.T}T))$, respectively. The treatment effect $\beta_{Y.T}$ was 0.5 when treatment had an effect and the outcome was continuous, 1 when treatment had an effect and the outcome was binary, and 0 otherwise. The other parameter values varied according to the scale of the covariates: For continuous covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{T.6} = \beta_{T.8} = \beta_{Y.1} = \beta_{Y.3} = \beta_{Y.5} = \beta_{Y.7} = 1$, $\beta_{T.1} = \beta_{T.3} = \beta_{T.5} = \beta_{T.7} = \beta_{Y.2} = \beta_{Y.4} = \beta_{Y.6} = \beta_{Y.8} = 0.05$, $\beta_0^T = 0$, and $\beta_0^Y = 0.5$. For mixed-scale covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{Y.1} = \beta_{Y.3} = 3$, $\beta_{T.6} = \beta_{T.8} = \beta_{Y.5} = \beta_{Y.7} = 1$, $\beta_{T.1} = \beta_{T.3} = \beta_{Y.2} = \beta_{Y.4} = 0.15$, $\beta_{T.5} = \beta_{T.7} = \beta_{Y.6} = \beta_{Y.8} = 0.05$, $\beta_0^T = -1.2$, and $\beta_0^Y = -0.7$. For binary covariates, $\beta_{T.2} = \beta_{T.4} = \beta_{T.6} = \beta_{T.8} = \beta_{Y.1} = \beta_{Y.3} = \beta_{Y.5} = \beta_{Y.7} = 3$, $\beta_{T.1} = \beta_{T.3} = \beta_{T.5} = \beta_{T.7} = \beta_{Y.2} = \beta_{Y.4} = \beta_{Y.6} = \beta_{Y.8} = 0.15$, $\beta_0^T = -2.5$, and $\beta_0^Y = -2$.