Received: 29 June 2018 | Revised: 29 January 2020 | Accepted: 3 February 2020

DOI: 10.1002/bimj.201800190

Biometrical Journal

RESEARCH PAPER

Automatic variable selection for exposure-driven propensity score matching with unmeasured confounders

Daniela Zöller[1,2] | **Leesa F. Wockner**[3] | **Harald Binder**[1,2]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

[2]Freiburg Center of Data Analysis and Modelling, Mathematical Institute – Faculty of Mathematics and Physics, University of Freiburg, Freiburg, Germany

[3]Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

Correspondence
Daniela Zöller, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany.
Email: zoeller@imbi.uni-freiburg.de

Funding information
Deutsche Krebshilfe; Federal Ministry of Education and Research, Grant/Award Number: 01GL1718A

Abstract

Multivariable model building for propensity score modeling approaches is challenging. A common propensity score approach is exposure-driven propensity score matching, where the best model selection strategy is still unclear. In particular, the situation may require variable selection, while it is still unclear if variables included in the propensity score should be associated with the exposure and the outcome, with either the exposure or the outcome, with at least the exposure or with at least the outcome. Unmeasured confounders, complex correlation structures, and non-normal covariate distributions further complicate matters. We consider the performance of different modeling strategies in a simulation design with a complex but realistic structure and effects on a binary outcome. We compare the strategies in terms of bias and variance in estimated marginal exposure effects. Considering the bias in estimated marginal exposure effects, the most reliable results for estimating the propensity score are obtained by selecting variables related to the exposure. On average this results in the least bias and does not greatly increase variances. Although our results cannot be generalized, this provides a counterexample to existing recommendations in the literature based on simple simulation settings. This highlights that recommendations obtained in simple simulation settings cannot always be generalized to more complex, but realistic settings and that more complex simulation studies are needed.

KEYWORDS

automated variable selection, exposure-driven matching, propensity score, unmeasured confounders

1 | INTRODUCTION

When analyzing follow-up data from epidemiological or clinical cancer registries, typically a comparison to an external reference population is desirable. For instance, having contacted a representative sample of cancer survivors from an epidemiological cancer registry (see, e.g., Beutel et al., 2015), one can then use general population reference values to evaluate late effects of cancer. Nevertheless, there will probably be baseline differences between the cancer survivor cohort and the reference cohort. The analysis might fail to capture the real causal late effects of cancer and cancer treatment. In addition, a high number of

available covariates allow for a more comprehensive characterization of the registry members. When judging the effect solely based on the general population reference values, these covariates and thus the information they contain is ignored.

The availability of large, general population-based cohorts provides an alternative for comparison. By identifying and matching similar individuals from the general cohort to the cohort of interest, one can estimate the long-term effect of having cancer and the potential harms of therapy. More technically, we consider membership in the cohort of interest as an exposure, and membership in the general cohort as non-exposure. We then apply exposure-driven matching techniques based on propensity scores to match individuals from the general population, that is, non-exposed individuals, to each cancer survivor. As described by Sjölander et al. (2012), the marginal effect estimated in the matched cohort can provide causal effect estimates for the exposed individuals, for example of cancer-related long-term health effects as compared to not having been exposed to cancer and the corresponding treatment. However, the challenges of propensity score modeling are different from those of standard regression modeling, as the aim of the models is different. For example, both Austin, Grootendorst, and Anderson (2007) and Weitzen, Lapane, Toledano, Hume, and Mor (2005) note that goodness-of-fit tests bear little relation to the ability of the estimated propensity scores to balance essential confounders across exposed and non-exposed individuals. There is an ongoing discussion regarding how to build the corresponding models in such an exposure-driven matching scenario.

When building the propensity score model, Patrick et al. (2011) suggest that selecting covariates associated strongly with the exposure, but unrelated to the outcome, should be avoided as this may increase bias. Brookhart et al. (2006) indicate that variables related to the outcome should always be included, regardless of their association with the exposure, which is supported by Rubin and Thomas (1996). In contrast, Austin (2007) found the bias to be lowest when including variables only associated with the exposure.

Post-matching, it is also unclear whether and to what extent further model adjustment is required to estimate the exposure effect on the outcome when focusing on settings with the marginal odds ratio as the estimand. Here, Sjölander and Greenland (2013) indicate that no further adjustment might be needed to obtain unbiased results, but if further adjustment is performed, one must always also adjust for the propensity score variables. Often this adjustment consists only of linear terms, despite the availability of more flexible adjustment strategies (Abrahamowicz, Schopflocher, Leffondré, du Berger, & Krewski, 2003; Benedetti and Abrahamowicz, 2004; Wynant & Abrahamowicz, 2014). As seen above, there are several recommendations in the literature for multivariable regression model building in the considered scenario. However, the recommendations were obtained in rather simple settings, and the question arises as to whether these recommendations can be generalized to real data situations. In reality, the observed data often combine a complex correlation structure, different covariate distributions, measurement errors, unmeasured confounders, and a limited number of observations. In contrast to this, many simulation studies use only a few variables. For example, Brookhart et al. (2006) only consider three variables, one of which relates to both the outcome and exposure, a second to the outcome but not the exposure, and the last only to the exposure. Rather simple settings are typically also considered concerning the joint distribution of covariates. For example, only binary or normally distributed variables are used in Austin (2007). Often the probability of exposure is generated as a function of the explanatory variables, rather than a random realization from a joint distribution (Austin, 2007; Wyss, Girman, LoCasale, Alan Brookhart & Stürmer, 2013). Stampf, Graf, Schmoor, and Schumacher (2010) present a more realistic scenario, with nine covariates (plus two interactions), some of which are correlated, from a mixture of distributions. Here, however, their primary aim was to demonstrate the counterfactual marginal odds ratio estimate rather than to investigate propensity score models.

In this study, we focus on a complex simulation design, which is based on a true observational study (Zöller, Wockner, & Binder, 2020). Here, it is difficult to reliably determine whether a covariate affects the exposure, the outcome of interest, or both, and consequently whether it should be included in the propensity score model or the outcome model. Although the true propensity score model in this design is difficult to determine and the results cannot be generalized, we can still add more evidence to the existing recommendations and potentially present a counterexample to them. As the true data generation process is often not fully known, we employ automatic variable selection strategies. We focus on backward selection strategies as recommended by Heinze, Wallisch, and Dunkler (2018). Additionally, in observational settings, unmeasured confounding is often present, and it is unclear how this might influence the performance of the considered strategies. In general, we do not try to differentiate between bias sources, but focus on the potential of different modeling strategies to correct bias independent of the source. To investigate this, some covariates will be explicitly removed from modeling in the simulation study to mimic unmeasured confounders, which have the potential to cause significant bias in estimates of exposure effects (Rubin, 1997). To this end, Section 2 introduces a simulation design mimicking several real-world challenges, and describes different selection strategies for the propensity score model and for the post-matching model, which we evaluate in this setting. We present an exemplary simulation-run analysis and the results of the complete simulation study in Section 3, in particular, investigating the effects of different unmeasured confounder patterns on variable selection. In Section 4, we discuss our result and offer concluding remarks. Additional simulation settings can be found in Supporting Information.

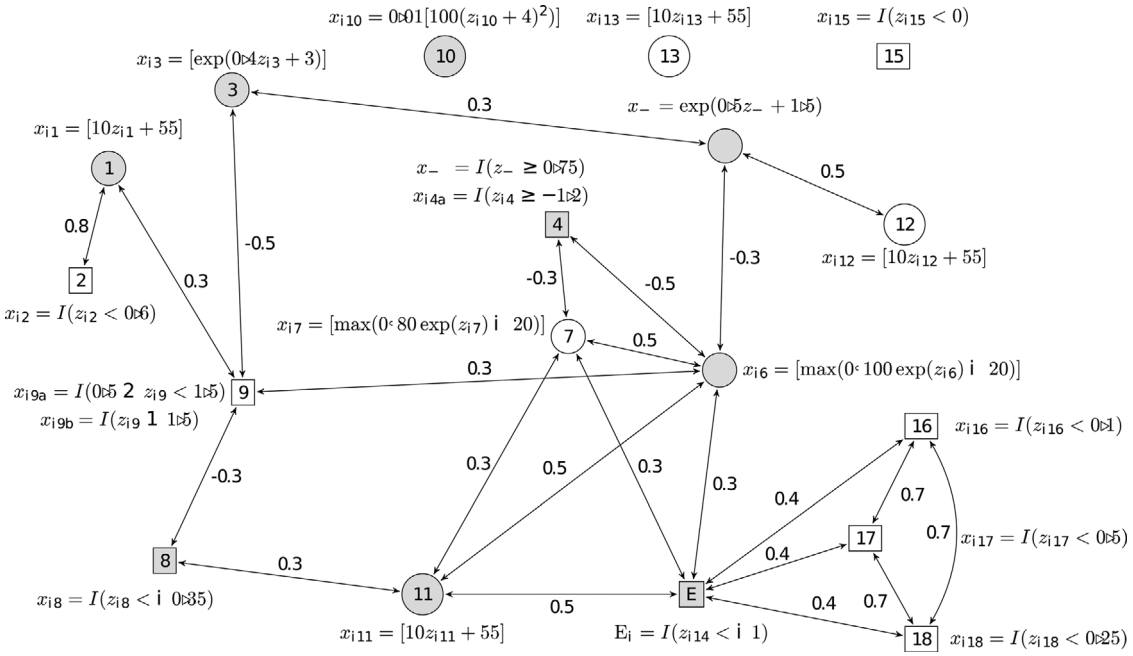

FIGURE 1 Simulation design: 17 z variables are generated following a multivariate normal distribution and later transformed. Arrows indicate the correlation structure and the numbers indicate the correlation coefficient of the multivariate standard normal distribution. If two variables are not connected via an arrow, the correlation coefficient is zero. The formulas for obtaining the covariates from the underlying variables are adjacent to the circles/rectangles. [.] indicates that we removed the non-integer part of the argument, and $I()$ is the indicator function, taking the value 1 if its argument is true and 0 otherwise. Circles indicate continuous variables, rectangles categorical variables, and E the exposure. If a covariate affects the response, the circle/rectangle is shaded. Variable 4 corresponds to both x_{4a} and x_{4b}, however only x_{4a} has a non zero effect on the outcome

2 | METHODS

2.1 | Simulation study

We consider a modified version of the "ART" study, which provides a simulation design for an artificial but realistic data set as described by Royston and Sauerbrei (2008). This simulation design is based on a large breast cancer data set (Sauerbrei & Royston, 1999; Schmoor, Olschewski, & Schumacher, 1996) and models the distribution of the predictors and their correlation structure. The 17 underlying variables (z) are drawn from a multivariate normal distribution with variance equal to one and mean zero, and are subsequently transformed into the actual covariates. The correlation structure and the transformations can be seen in Figure 1. Some variables are transformed, truncated, or have the non-integer portion of the variable removed. Such transformations mimic both complex real distributions as well as measurement errors of observational data. The simulation design has also been used by Binder, Sauerbrei, and Royston (2013) to compare multivariate model building strategies, albeit with nonlinear effects. Binder, Sauerbrei, and Royston (2011) presented further details about the design of this simulation as a technical report. The outcome y_i, $i = 1, \ldots, n$, for each individual is randomly generated from a Bernoulli distribution with mean p_i such that,

$$p_i = \frac{1}{1 + \exp(-\eta_i)},$$

where

$$\eta_i = \sigma^{-1}(\omega + \lambda \mathrm{E}_i - 0.015x_{i1} + 0.03x_{i3} - 0.4x_{i4a} - 0.12x_{i5} + 0.018x_{i6} + 0.4x_{i8} + 0.021x_{i10} + 0.04x_{i11}). \quad (1)$$

In the original simulation design by Binder et al. (2011), there were nonlinear effects on η_i. We replaced these nonlinear effects by linear effects such that the explained variance measured by R^2 is similar to the original. The code for the modified simulation design is published on Zenodoe (Zöller et al., 2020).

Based on the above-described simulation design, we considered two different scenarios with the total sample size equal to 2,500. Across 10,000 simulation runs, the average number of exposed individuals was 397. We designed both scenarios such that

TABLE 1 A summary of the two simulation scenarios

Scenario	R^2	Marginal OR	$P(Y = 1)$	ω	λ	σ
SNR = 0.1	0.09	2	50%	−3.6123	0.8544	1.1646
SNR = 0.4	0.29	2	50%	−3.5188	0.4124	0.4985

the true marginal odds ratio, as suggested by Austin and Stafford (2008), is about 2 and the proportion of individuals with the outcome ($Y = 1$) is about 50%. The signal-to-noise ratio (SNR) introduces the only difference in the two scenarios: In the first scenario, the SNR is designed to be about 0.1 and in the second about 0.4. These signal-to-noise ratios correspond to correctly specified models with R^2 values of approximately 0.09 and 0.29, respectively. We achieve these two settings by choosing three parameters: The offset ω, the exposure effect λ on the linear predictor, and the scaling parameter σ. In the first setting, the following values were chosen: $\omega = -3.6123$, $\lambda = 0.8544$ and $\sigma = 1.1646$. In the second setting, the values were chosen to be $\omega = -3.5188$, $\lambda = 0.4124$ and $\sigma = 0.4985$. Due to the complex simulation design, the true marginal effect cannot be obtained analytically. Instead, we calculate the probability of experiencing the outcome for exposed individuals both under the condition that they are exposed and under the counterfactual condition of being unexposed based on Equation (1). Subsequently, we calculate the marginal odds ratio. A summary of the two simulation scenarios is shown in Table 1.

The simulation design mimics, for example, a case study in which the outcome Y can be seen as 5-year progression-free survival and the exposure as an indicator for whether or not a breast cancer patient was treated with radiotherapy according to national guidelines. The design attempts to capture the complexities of real data. For example, the variable x_6 is based on progesterone receptor status, a typical clinical variable measured at baseline. This is highly correlated to estrogen receptor status (x_7, $\rho = 0.5$) and tumor grade 3 versus 1,2 (x_4, $\rho = -0.5$). Both progesterone receptor status and estrogen receptor status are correlated to the exposure, but only progesterone receptor status is related to the outcome. This connection between the variables suggests that if x_6 is an unobserved covariate, x_7 can act as a proxy for x_6 in modeling the exposure. That is, x_6 is a variable whose effect can mostly be approximated by other variables related only to the exposure. If x_6 is unavailable for selection, we would expect a strategy that only includes variables related to both the exposure and the outcome to perform poorly as it would fail to identify the confounding effect of x_6 and not consider x_7 as a replacement.

The variable x_{11} (a variable based on age) is similar to x_6 in that a combination of other variables can approximate its effect. However, unlike x_6, it can be replaced with variables related to both the exposure and outcome (namely x_6). Thus in this scenario, we would expect a strategy that included all variables associated with the exposure in the propensity score model to have an unnecessarily large variance estimate.

The variable x_8 is a binary variable purporting to indicate whether or not a patient received Tamoxifen treatment. Although not strictly a baseline variable, it is often considered as a prognostic factor. The variable x_8 has a surprising influence on both the outcome and the exposure. Although x_8 is not directly related to the exposure, in the presence of x_{11} it is associated with the exposure. This influence is observable as x_{11} is a continuous variable modeling a binary outcome. The inclusion of x_8 allows for a more accurate estimate of the x_{11} effect. Hence, when x_8 is unavailable, we are likely to poorly estimate the confounding effect of x_{11} and hence the propensity score, and also be unable to accurately estimate the age effect in the outcome model.

In each simulation scenario, four simulation settings were considered using the same data. To mimic the effect of unmeasured covariates, a single variable was unavailable for selection. The first two simulation settings (A-1 and A-2) represent situations, where the unmeasured variable is either irrelevant (A-1: x_{15} omitted) or can be approximated rather easily with variables related to the exposure but not the outcome (A-2: x_6). The other two simulation settings (B-1 and B-2) represent a more difficult situation. In setting B-1, x_{11} is omitted and thus the unmeasured variable can be approximated best with a variable related to both the outcome and the exposure, which makes the estimation of both the effect of x_6 as well as x_{11} difficult. In setting B-2, x_8 is omitted and the situation is comparable to B-1, with the exception that x_8 is per definition no confounder itself but is correlated with one. See Table 2 for a summary of the simulation scenarios and the omitted variables. The simulations were repeated $M = 10\,000$ times. The results for the simulation settings concerning x_6, and x_8 are presented in the following sections. Please refer to Supporting Information for the results concerning x_{11}, and x_{15}.

2.2 | Propensity score model selection

The propensity score (PS) model is estimated using a logistic regression model for the exposure. The variables included in this model are selected using one of four criteria: variables are included which are deemed to be associated with the exposure (E), the outcome (O), both the exposure and the outcome ($E - O$), or either the exposure or the outcome (E/O). The selection itself is performed using backward selection in a logistic regression model for the exposure or the outcome, respectively, either based

TABLE 2 Simulation settings and characteristics of omitted variables

Simulation setting	Example	Confounder	Correlated with other variables	Can be approximated by variables related to:		
				Exposure	Outcome	Exposure and Outcome
A-1	x_{15}	No	No	✗	✗	✗
A-2	x_6	Yes	Yes	✓	✗	✗
B-1	x_{11}	Yes	Yes	✓	✓	✓
B-2	x_8	No	Yes	✗	✗	✗

on the p-value with the cutoff α set to 0.05, or AIC. The $E - O$-strategy is achieved by including only the variables which are selected in a logistic regression for the outcome and for the exposure, respectively, whereas the E/O-strategy combines the two variable sets in the propensity score model. The scenarios resulting from the combination of the model selection criteria and the backward selection method are denoted E(AIC), E(0.05), O(AIC), O(0.05), $E - O$(AIC), $E - O$(0.05), E/O(AIC), and E/O(0.05). As a comparison, we include a naive propensity score model where only variables with a nonzero correlation with the exposure in the underlying generation process are included, namely $x_{6,\,7}$, x_{11}, x_{16}, x_{17}, and x_{18} if available for selection.

2.3 | Matching via propensity score

Investigation of propensity score matching methods suggests that nearest neighbor matching (in random order or closest distance) on the logit of the propensity score with a caliper tends to result in slightly less biased estimates than other matching algorithms (Austin 2014, 2017). Cochran and Rubin (1973) demonstrate that a caliper defined to have a maximum width of 0.2 standard deviations of the estimated propensity score removes approximately 99% of the bias due to normally distributed confounding while estimating a linear exposure effect. This caliper size is further supported by Austin (2011) based on a series of Monte Carlo simulations.

As recommended, we match the observations on the estimated logit propensity score, and a caliper of 0.2 is set such that for each exposed individual, we choose a non-exposed individual with the closest logit propensity score within 0.2 standard deviations as its matching partner. We discard observations without a match within 0.2 standard deviations of the logit propensity score. If a specific variable selection strategy does not include any variables in the propensity score model, every individual is predicted to have the same propensity score, and we randomly match non-exposed individuals to exposed individuals. While this procedure results in a different proportion of exposed individuals being matched to non-exposed individuals, Austin et al. (2007) indicate that differences in the proportion of matched pairs does not necessarily result in a difference in the balance of the matched data set. R version 3.4.0 (R Core Team, 2016) and the package `MatchIt` version 3.0.1 (Ho, Imai, King, & Stuart, 2011) was used to facilitate 1-1 matching in a random order.

2.4 | Post-matching adjustment

After matching, the exposure and outcome are related using logistic regression for the outcome with exposure as a covariate, and we investigate different model-building strategies. First, assuming there are no additional confounders, we do not adjust for further variables using the exposure-discordant odds ratio described below (denoted by *ED OR*). However, in applications, there are often additional confounders as in our simulation design. Consequently, we consider the option of additional adjustment in the post-matching outcome model. Specifically, we investigate several post-matching adjustment strategies which include additional variables in the model. Under the assumption that additional adjustment is needed in the outcome model, Sjölander and Greenland (2013) demonstrate that one also needs to adjust for variables in the propensity score model. Accordingly, we only consider strategies that adjust at least for these variables. First, we adjust for all available covariates ($All + E$). Next, we adjust only for variables in the propensity score model ($PSV + E$). Finally, we use backward selection and select variables for post-matching adjustment based on their association with the outcome in a model where the propensity score variables are mandatory. Again, we consider two selection thresholds, p-value ≥ 0.05 ($BS(0.05) + E$) and the AIC criterion (BS(AIC) + E).

2.5 | Estimation of marginal effects

The aim of our analysis is to estimate the causal effect the exposure had on the outcome of the individuals which were observed to be exposed. In accordance with Sjölander and Greenland (2013), we use the marginal effect estimated in the propensity score matched cohort by trying to find a proxy for the counterfactual outcome among the exposed individuals. Typically, the estimate of the exposure effect is only a conditional estimate. For example, results obtained from a logistic regression with

several covariates are conditional odds ratios, which might differ from the marginal odds ratio. One way to obtain a marginal odds ratio is to estimate the exposure-discordant odds ratio (δ_{ED}) as shown by Sjölander et al. (2012). Given n pairs of matched observations consisting of an exposed and non-exposed individual, we can summarize the matched pairs as such: Let n_{00} indicate the number of matched pairs who both do not experience the event and n_{11} the number of pairs where both individuals experience the event. Let n_{01} indicate the number of pairs where the exposed individual experienced the event and n_{10} the number of pairs where the non-exposed individual experienced the event. The exposure discordant odds ratio is then given by

$$\delta_{\text{ED}} = \left(\frac{(n_{11} + n_{10})(n_{10} + n_{00})}{(n_{11} + n_{01})(n_{01} + n_{00})} \right).$$

The variance for this estimate is further detailed in Sjölander et al. (2012). Nevertheless, no post-matching adjustment can be incorporated.

Alternatively, a counterfactual marginal odds ratio estimator first proposed by Graf and Schumacher (2008) was used. It aims to estimate the ratio of the odds if everybody versus nobody was exposed taking confounders into account. Consider the tuple (Y, X, E), where X is a vector of covariates, E is a binary exposure indicator, and Y is the (binary) response variable. Let the marginal probabilities of the outcome be given by p_0 and p_1. Conditional on X, let the probability of the outcome be given by

$$p_1(X) = P(Y = 1|X, E = 1) \text{ and } p_0(X) = P(Y = 1|X, E = 0),$$

depending on whether $E = 1$ or $E = 0$, respectively. If the assumption of "strongly ignorable treatment assignment" (SITA) holds, then

$$\begin{aligned} E_X\big(p_e(X)\big) &= E_X P(Y = 1|X, E = e,), \\ &= E_X P(Y_e = 1|X), \\ &= p_e, \end{aligned}$$

where E_X is the expectation with respect to the distribution of X, for $e = 0, 1$, and Y_1 is the outcome for an exposed individual and Y_0 the outcome for the same individual, but unexposed (Rosenbaum & Rubin, 1983). The conditional odds ratio under SITA, given covariates X, is thus derived by contrasting the conditional response probabilities with and without treatment. Similarly, the marginal OR, the ratio of the odds if everybody versus nobody was exposed, can be defined by

$$\delta_{\text{CF}} = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)}. \tag{2}$$

Graf and Schumacher (2008) further proposed two procedures to estimate p_1 and p_0: one stratifying by propensity score, analogous to the Mantel–Haenzel estimator; the other using a logistic model based on data matched via the propensity score and potential post-matching adjustment. We employ the latter procedure to estimate p_0 and p_1 when considering post-matching adjustment. Consider an independent random sample (y, x, e) of size n and corresponding design matrix $\Pi = (\mathbf{1}, x, e)$ with dimension $(n \times p)$. In the following, we will use a logistic regression model:

$$\text{logit}\{P(y = 1|x, e)\} = \beta_0 + x\beta_x + e\beta_e.$$

The maximum likelihood estimators $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta'}_x, \hat{\beta'}_e\}$ was used to estimate the marginal probabilities of response based on individual probabilities of response, $\hat{p}^i_{e,LR}$, $e = 1, 0$, estimated from the logistic regression:

$$\hat{p}_{e,LR} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}^i_{e,LR} \text{ with } \hat{p}^i_{e,LR} = (1 + \exp\{-(\hat{\beta}_0 + \hat{\beta}_{e'} e + x'_i \hat{\beta}_x)\})^{-1}. \tag{3}$$

Substituting Equation (3) into Equation (2), the marginal effect of E is obtained. The variance estimate of this estimator is further derived in Stampf et al. (2010).

2.6 | Performance measures

Each propensity score selection strategy may result in different propensity score models, while the models may also vary across simulation runs. To give an impression of the most frequent models, we visualize these models. The four most frequent models

are compiled separately for each simulation scenario, simulation setting, and each propensity score selection strategy. We then plot each unique model with a matrix plot where each row represents a distinct model and each column a covariate of the simulation design. If a covariate is selected into the model of interest, the corresponding box is colored light grey. If a covariate is not selected, the corresponding box is colored black. Additionally, the frequency of each model in each scenario across all simulation runs is plotted using a heatmap where each row also represents a distinct model with the same order and number as in the matrix plot. A lighter gray indicates a higher model frequency, while dark gray indicates a lower model frequency.

The bias of the marginal odds ratio on the log scale is calculated as $\mathrm{E}[\log(\hat{\theta})] - \log(\theta)$, where $\hat{\theta}$ denotes the estimated marginal odds ratio derived by either the exposure-discordant odds ratio (δ_{ED}), or the adjusted counterfactual marginal odds ratio estimater (δ_{CF}), and $\log(\theta)$ denotes the true log marginal odds ratio.

We present the average standard deviation of the log marginal odds ratio of exposure in the outcome model as a measure of variability and calculate the total number of variables used in the modeling process. To measure the amount of information remaining in the matched data set, we measure the average number of matched individuals in the final data set. Furthermore, we analyze the average number of variables used in the complete analysis. We argue that when there is no difference in the bias of the log marginal odds ratio, a smaller model should be favored as in reality a limited number of observations also limits the number of variables which can be included in the models. In particular, we argue that a full model using all variables is not possible in most practical situations.

3 | RESULTS

3.1 | Exemplary analysis of one simulation run

We illustrate the different model building strategies for the propensity score and the outcome model post-matching using the data from one exemplary simulation run in the setting with a signal-to-noise ratio of 0.4. In this specific run, the variable x_6 is not available for the analysis (simulation setting A-2). Note that the results presented in this section are only given to demonstrate the procedures themselves, but cannot be generalized.

Although the simulation is based on a pre-defined correlation structure shown in Figure 1, the actual data set is obtained via subsequent transformations, and thus the true correlation structure is unknown. As an estimate, we present the empirical correlation structure in Figure 2. We used the Pearson correlation to measure the association between two continuous variables and between a continuous and a binary variable, and the absolute of the logarithm of the odds ratio for two binary variables. The latter is scaled with the highest observed logarithmic odds ratio ($\max(|\log(\theta)|) = 5.27$). In general, the observed values are slightly smaller than the nominal ones, but the basic structure is maintained. In particular, one can see that many variables are correlated or connected through different variables. For example, x_{4a} is correlated with x_6 but not x_{11}, but x_6 itself is correlated with x_{11}. Such a correlation structure challenges any variable selection procedure.

Due to the data structure, there is a large difference between the covariate distributions of the exposed and the unexposed individuals. We measured the difference by calculating the mean covariate differences between the two populations (Table 3). We observe differences for example in x_6 and x_{11} (However, note the different scale of these two variables), which both have an effect on the outcome and fulfil the definition of a confounder. Thus, these differences are problematic when estimating the causal effect of the exposure. The matching of an unexposed individual to every exposed individual based on the propensity score aims to balance out these differences. Specifically, we aim to obtain the same covariate distribution in the matched cohort as in the exposed population.

The selected variables for the propensity score are printed in bold in Table 3. As x_6 is assumed to be unmeasured, the naive propensity score model (NaivePS) uses the variables x_7, x_{11}, x_{16}, x_{17}, and x_{18}. Based on the different estimations of the propensity score, we obtain different sample sizes in the post-matching analysis cohort. The number of exposed individuals is 402 in the specific simulation run, meaning that even if we were able to find a suitable matching partner for every exposed individual, the analysis data set would consist of a maximum of 804 individuals and 1,696 of the 2,500 individuals would be discarded by default. When using the naive propensity score model, we only find a matching partner for 379 of the 402 exposed individuals and have to discard the rest. On average, the difference between the exposed and unexposed individuals is drastically reduced, but we still observe a difference in the unmeasured variable x_6 and the difference between some variables, such as x_{12}, is increased.

The smallest propensity score model was achieved when requiring an association with both the exposure and the outcome and using a p-value with a cutoff of 0.05 ($E - O(0.05)$ as a selection criterion. Here, x_{4a}, x_7, x_8, and x_{11} are selected and we were able to identify a matching partner for 399 of the 402 exposed individuals. All selected variables are correlated with the

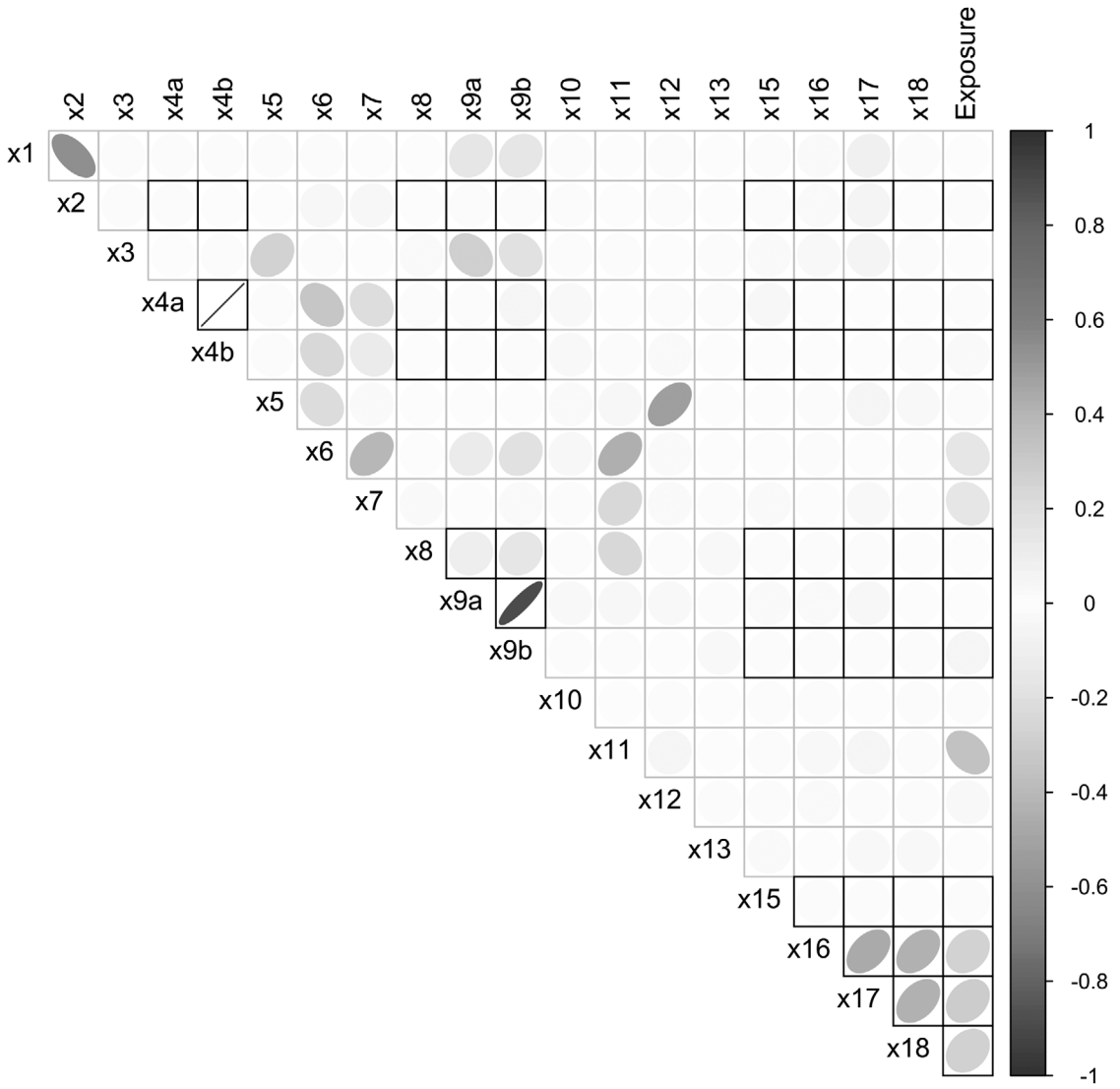

FIGURE 2 Visualization of the association between the variables in one simulation run. The association between two continuous and between a continuous and a binary variable is measured via Pearson correlation coefficient and the association between two binary variables via the logarithm of the odds ratios scaled with the highest observed logarithmic odds ratio ($\max(|\log(\theta)|) = 5.27$). The associations measured via logarithmic odds ratio are indicated with darkened boxes. The stronger the association, the darker the color and the closer the ovals are to lines. A positive correlation is indicated by an orientation of the oval from lower left to upper right, whereas a negative correlation is indicated by an orientation of the oval from upper left to lower right

unmeasured confounder and one might hope that these have the potential to compensate for some of the missing information. However, the difference in x_6 is greater than when using the naive propensity score model, while the differences are on average strongly decreased. If one uses the AIC instead of the p-value ($E - O(\text{AIC})$), x_{12}, x_{16}, and x_{17} are additionally selected, and 30 additional individuals are discarded. On average, the smallest difference between the exposed and unexposed individuals is obtained, but we still observe a difference in some variables such as x_6 and x_7 while simultaneously reducing the sample size the most. Thus, the later estimated effect of the exposure might only be valid for a subset of the exposed individuals and not for the ones similar to the discarded individuals, which has to be discussed in a real application. Requiring an association with only the exposure (E), only the outcome (O), or with either the exposure or the outcome (E/O) increases the number of variables in the propensity score models and roughly speaking also decreases the difference between the populations, but only to a small degree overall. The example demonstrates that it is very difficult to choose the correct propensity score model and to remove all differences between the two populations.

The aim of the following analysis is to estimate the causal effect of the exposure on the outcome of the exposed individuals. For this purpose, we calculate the marginal odds ratio using the propensity score matched data sets, which we simulated to

TABLE 3 Number of individuals used (n) and mean covariate values of exposed individuals minus mean values of unexposed individuals for the original data (Orig.) and after matching based on the propensity score with different propensity score models in one simulation run where x_6 (marked in italics) was assumed to be unmeasured and simulation scenario 2 was used

	Orig.	Naive PS	E(AIC)	E(0.05)	O(AIC)	O(0.05)	$E-O$(AIC)	$E-O$(0.05)	E/O(AIC)	E/O(0.05)
n	2500	758	752	758	768	798	768	798	742	738
x_1	0.62	1.42	0.26	1.34	**0.09**	**0.41**	1.29	0.89	**0.16**	**0.27**
x_2	0.00	−0.03	0.00	−0.02	**0.02**	**−0.01**	−0.03	−0.03	**−0.02**	**−0.02**
x_3	0.37	0.10	0.48	0.67	**0.54**	**−0.13**	−0.06	0.83	**0.36**	**0.18**
x_{4a}	−0.03	−0.03	**0.01**	**0.00**	**−0.02**	**0.01**	**−0.01**	**−0.02**	**0.00**	**0.01**
x_{4b}	−0.01	−0.05	**−0.01**	−0.02	−0,03	−0.04	−0.05	−0.04	**0.01**	−0.03
x_5	0.13	0.25	0.18	0.38	0.11	**−0.03**	−0.04	0.14	**0.16**	**0.16**
x_6	*−75.47*	*1.26*	*−6.18*	*−9.12*	*−11.22*	*−7.73*	*−1.14*	*2.28*	*−5.73*	*−10.47*
x_7	−65.79	**−0.72**	**2.24**	**−3.44**	**−1.26**	**−1.78**	**0.47**	**0.44**	**−0.68**	**−3.30**
x_8	0.01	−0.09	**0.04**	**0.01**	**0.02**	**0.00**	**0.03**	**−0.02**	**0.00**	**−0.01**
x_{9a}	0.01	0.04	0.02	0.04	−0.01	**−0.01**	0.01	0.02	**−0.02**	**0.01**
x_{9b}	0.01	0.00	0.02	0.02	0.01	**0.02**	0.02	−0.01	**−0.01**	**0.01**
x_{10}	0.01	−0.29	−0.25	0.18	−0.46	**−0.58**	0.18	0.38	**−0.30**	**−0.11**
x_{11}	−9.01	**−0.23**	**−0.78**	**−0.47**	**−0.38**	**−0.32**	**0.01**	**−0.13**	**0.13**	**−0.66**
x_{12}	0.56	1.87	**0.34**	**1.42**	−0.60	0.76	**−0.45**	1.07	**0.48**	0.44
x_{13}	0.30	−0.15	0.68	0.70	1.08	−0.11	0.39	0.87	0.35	−0.35
x_{15}	0.00	−0.04	−0.03	−0.02	−0.04	−0.01	0.00	0.02	−0.03	**0.00**
x_{16}	0.29	**0.00**	**0.01**	**0.00**	**0.00**	0.34	**0.03**	0.35	**0.01**	**−0.02**
x_{17}	0.26	**0.00**	**−0.01**	**0.01**	0.15	0.31	0.20	0.30	**0.02**	**−0.01**
x_{18}	0.27	**0.02**	**0.00**	**0.01**	0.14	0.31	0.20	0.29	**0.03**	**−0.03**

Note. Naive PS: Naive propensity score model chosen to include x_7, x_{11}, x_{16}, and x_{17}. The other propensity score models were achieved via backward selection with either the AIC or the p-value with a cutoff of 0.05 used as the selection criterion. Association was required with either only the exposure (E), only the outcome (O), the exposure and the outcome ($E-O$), or either the exposure or the outcome (E/O). Variables included in the corresponding propensity score models are marked bold. The gray scaling is used to highlight variables which differ between the exposure groups.

be 2. The results for different estimating procedures are given in Table 4 including the standard deviation of the marginal odds ratio. Without additional adjustment (*ED OR*), the least biased results are obtained when building the propensity score model using $E-O$(AIC). As seen above, the corresponding propensity score model achieved the best balance over all considered strategies. However, the result is still somewhat biased and can be improved by adjusting for covariates in the outcome model. Here, backward selection, independent of the selection criterion, has the lowest observed bias when combined with O(0.05), $E-O$(AIC), or $E-O$(0.05) for building the propensity score model. A comparable result is obtained when using the naive propensity score model and adjusting for all variables in the outcome model. The question here is whether these results, which are in line with the recommendations in the literature, are true on average and if other variables are unmeasured.

3.2 | Simulation results

The results of the visualization of the four most frequent propensity score models across the two unmeasured confounder scenarios x_6, and x_8 can be found in Figure 3 (for AIC as the selection criterion) and in Figure 4 (for p-value ≤ 0.05 as the selection criterion). When we use the AIC, we obtain 45 unique propensity score models, while we obtain 49 unique models when we apply the p-value ≤ 0.05 threshold. On the left side of the figures, the variables included in the model are shown. The arbitrary model number is displayed on the y-axis and the variable name on the x-axis, where a light-colored field means that the variable is included in the model. For example in Figure 3 model number 1, only the variable x_{11} is included, whereas in model 2 x_7, and x_8 are additionally included. On the right side of the figures, the frequency of the models shown on the left is displayed across all simulation runs separately for the variable removal settings and the propensity score modeling strategies. For example, propensity score model 18 comprising all variables except x_2, x_6, x_{13} ,and x_{15} is chosen in about 16% of all simulations when x_6 in unavailable for selection and the E(AIC)-strategy is used, whereas this model is not chosen in any simulation run when using the O(AIC)-strategy.

TABLE 4 Estimated marginal odds ratio and standard deviation of the marginal odds ratio in brackets for one simulation run after matching based on the propensity score with different propensity score models in one simulation run where x_6 was assumed to be unmeasured and simulation scenario 2 was used

	Unadjusted	Adjusted(+ Exposure + *PS* variables)			
	ED OR	***PS* variables**	***BS*(AIC)**	***BS*(0.05)**	**All variables**
Naive PS	1.60 (0.31)	1.68 (0.30)	1.87 (0.27)	1.87 (0.27)	**2.00 (0.27)**
E(AIC)	1.43 (0.30)	1.47 (0.29)	1.49 (0.26)	1.49 (0.26)	1.49 (0.26)
E(0.05)	1.70 (0.33)	1.94 (0.30)	2.27 (0.27)	2.27 (0.27)	2.37 (0.27)
O(AIC)	1.70 (0.32)	1.65 (0.27)	1.67 (0.27)	1.67 (0.27)	1.82 (0.28)
O(0.05)	1.70 (0.30)	1.90 (0.27)	**2.02 (0.28)**	**2.02 (0.28)**	1.89 (0.29)
$E-O$(AIC)	1.93 (0.34)	1.83 (0.32)	**2.02 (0.29)**	**2.02 (0.29)**	2.42 (0.30)
$E-O$(0.05)	2.24 (0.34)	2.18 (0.33)	**2.01 (0.30)**	**2.00 (0.30)**	1.90 (0.32)
E/*O*(AIC)	1.81 (0.33)	1.64 (0.28)	1.64 (0.28)	1.64 (0.28)	1.64 (0.28)
E/*O*(0.05)	1.70 (0.31)	1.91 (0.27)	1.87 (0.27)	1.87 (0.27)	1.85 (0.28)

The true marginal odds ratio is approximately 2, and the closest estimates are marked bold. The different methods for the propensity score model are displayed in the rows, including Naive PS (chosen to include x_7, x_{11}, x_{16}, and x_{17}) and eight different backward selection methods with either the AIC or the *p*-value with a cutoff of 0.05 used as the selection criterion. Association was required either with only the exposure (*E*), only the outcome (*O*), the exposure and the outcome ($E-O$), or either the exposure or the outcome (*E*/*O*). The marginal odds ratio was calculated using the exposure-discordant odds ratio (*ED OR*) or via adjusted logistic regression outcome models. For the latter, the models were adjusted for the propensity score variables (*PS* variables), with additional variables selected via backward selection with the AIC (*BS*(AIC)) or the *p*-value (*BS*(0.05)) as the selection criterion, or with all variables (All variables).

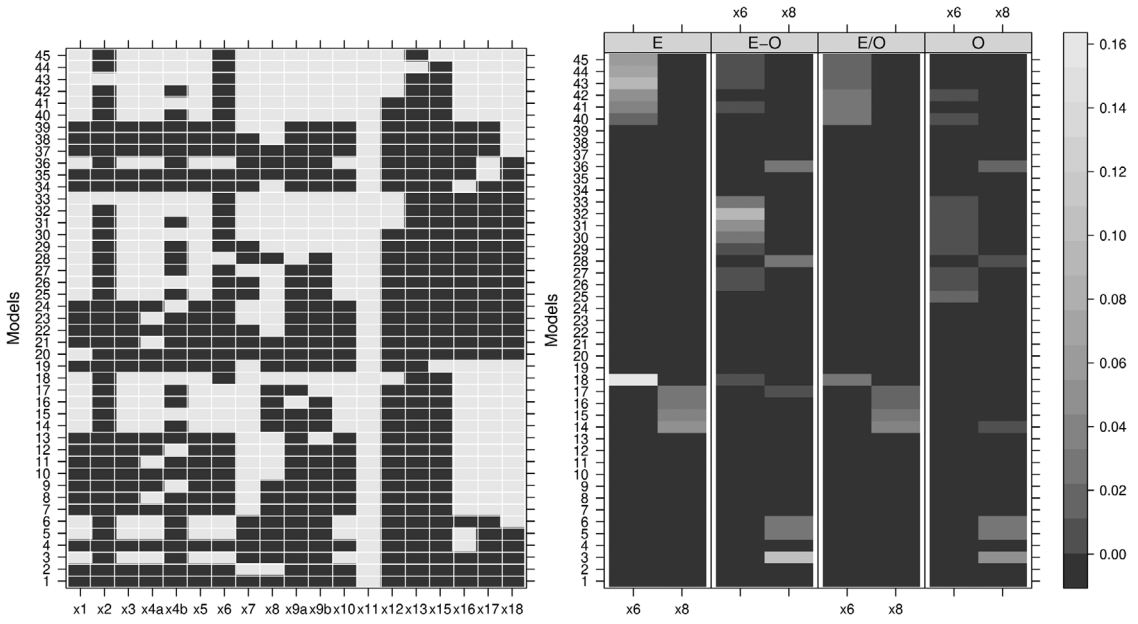

FIGURE 3 (left) The composition of the 45 most common propensity score models for all scenarios when the AIC threshold is used to select the propensity score. Each row represents a unique model, while the *x*-axis indicates the candidate variables. Lighter gray indicates that the variable was included in the model, while dark gray indicates that the variable was not included in the model. (right) Frequency of inclusion for the 45 most common models when the AIC threshold is used. The different scenarios are on the *x*-axis where x_6, and x_8 indicate the variable is unavailable for selection. Lighter gray indicates higher model frequency, while darker gray indicates lower model frequency

The removal of x_6, and x_8 is intended to represent unmeasured variables common in real world data corresponding to different scenarios. The average bias and the average standard deviation of the marginal log exposure effect, and the average number of variables used in the whole process are presented in Figure 5 with removal of x_6, and in Figure 6 with removal of x_8.

Across all removal of covariate settings, the E/O(AIC) strategy leads to the largest models, followed by E(AIC), O(AIC), and finally $E-O$(AIC) (see Figures 5 and 6, second panel). Next, the methods with the *p*-value as a selection criterion follow in the same order. This difference in the model sizes reflects that the $E-O$-strategy is more restrictive than the other strategies and the *p*-value-strategy is more restrictive than the AIC-strategy. In Table 5 the average number of matched individuals in the

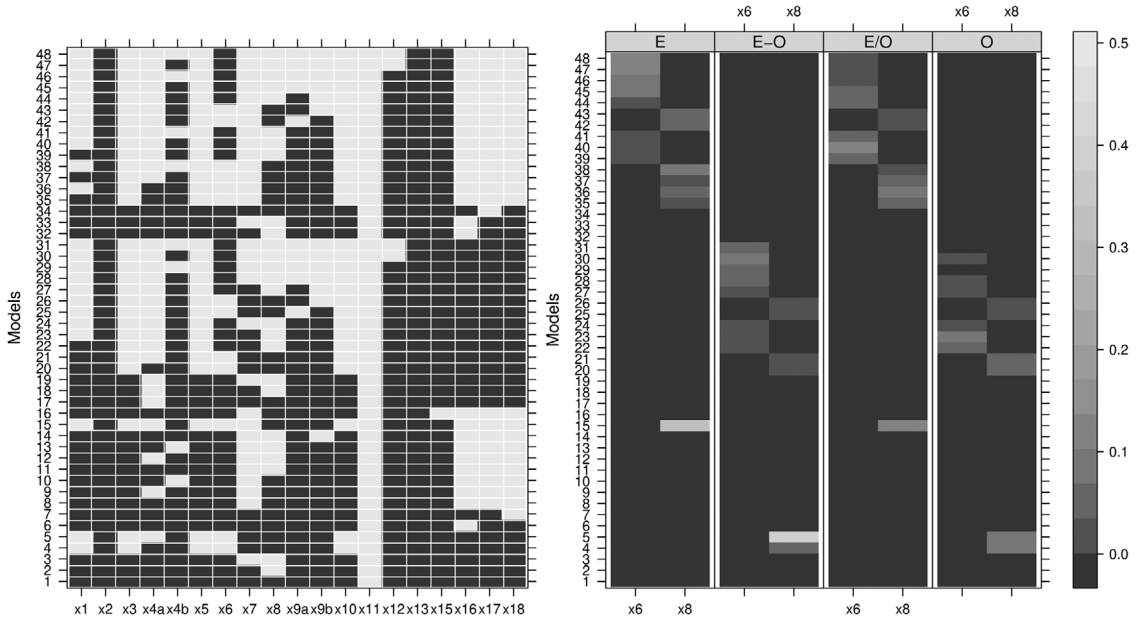

FIGURE 4 (left) The composition of the 48 most common propensity score models for all scenarios when the p-value ≤ 0.05 threshold is used to select the propensity score. Each row represents a unique model, while the x-axis indicates the candidate variables. Lighter gray indicates that the variable was included in the model, while dark gray indicates that the variable was not included in the model. (right) Frequency of inclusion for the 48 most common models when p-value ≤ 0.05 threshold is used. The different scenarios are on the x-axis where x_6, and x_8 indicate the variable is unavailable for selection. Lighter gray indicates higher model frequency, while darker gray indicates lower model frequency

final data set is displayed separately for all simulation settings, removal settings, and selection strategies. Given a simulation setting and removal setting, we observe smaller models with a higher average number of matched individuals.

When the Both Exposure and Outcome strategy ($E-O$) is used to build the propensity score model, there might be no overlap between the two variable sets selected in the logistic regression models for the outcome and the exposure, which would mean that no variable is included in the propensity score model. In this case, the estimated propensity score is the same for every individual, and we randomly match the exposed and the unexposed. In our simulation setting, this occurred very rarely (never in A-2 and only rarely in B-2, namely only in 0.25% with SNR = 0.1 and E − O(AIC), 3.62% with SNR = 0.1 and E − O(0.05), and 0.01% with SNR = 0.4 and E − O(0.05)) and it did not strongly influence the results. Nevertheless, this can also occur in practice and has to be kept in mind. It did not happen in our simulations for any of the other strategies.

3.2.1 | Simulation settings A

In the following, the simulation results of the simulation setting A-2 (Removal of x_6) are given. The results for the simulation setting A-1 can be found in Supporting Information and are roughly similar to the ones of A-2. If the covariate x_6 is unavailable for selection, the most frequent propensity score models included nearly all variables when association with the exposure is required (E(AIC) and $E(0.05)$) or sufficient (E/O(AIC) and $E/O(0.05)$). Only variables associated with neither the exposure nor the outcome (x_{13} and x_{15}), and x_2 were mostly not selected. This demonstrates the central role x_6 plays in differentiating between exposed and unexposed individuals. Due to their correlation with x_6, variables only indirectly correlated with the exposure are also selected as they carry some of the missing information of x_6. These variables would not be selected if x_6 were available (see Appendix B.1 in Supporting Information). When additionally requiring an association with the outcome ($E-O$(AIC) and $E-O(0.05)$), the models become a little smaller on average. Here, the variables x_{16} to x_{18} are additionally excluded as they are only indirectly correlated with the variables through the exposure. Focusing on an association with the outcome (O(AIC) and $O(0.05)$), the model size is only slightly reduced. As x_6 has a strong effect on the outcome, variables correlated with the unmeasured x_6 are included as a replacement in the model for the exposure. When using the p-value as the selection criterion rather than the AIC, variables with a lower (indirect) correlation with x_6 are excluded more often.

Comparing the number of propensity score matched individuals for the different strategies in Table 5, we observe that the largest number are matched when using the $E-O(0.05)$, or $O(0.05)$ strategies. The propensity score models where an association with the exposure is required lead to the smallest final sample sizes. Thus, in these cases it is more difficult to find similar matching partners for every exposed individual. As the propensity score models are large, it is more difficult to balance out all of these variables simultaneously.

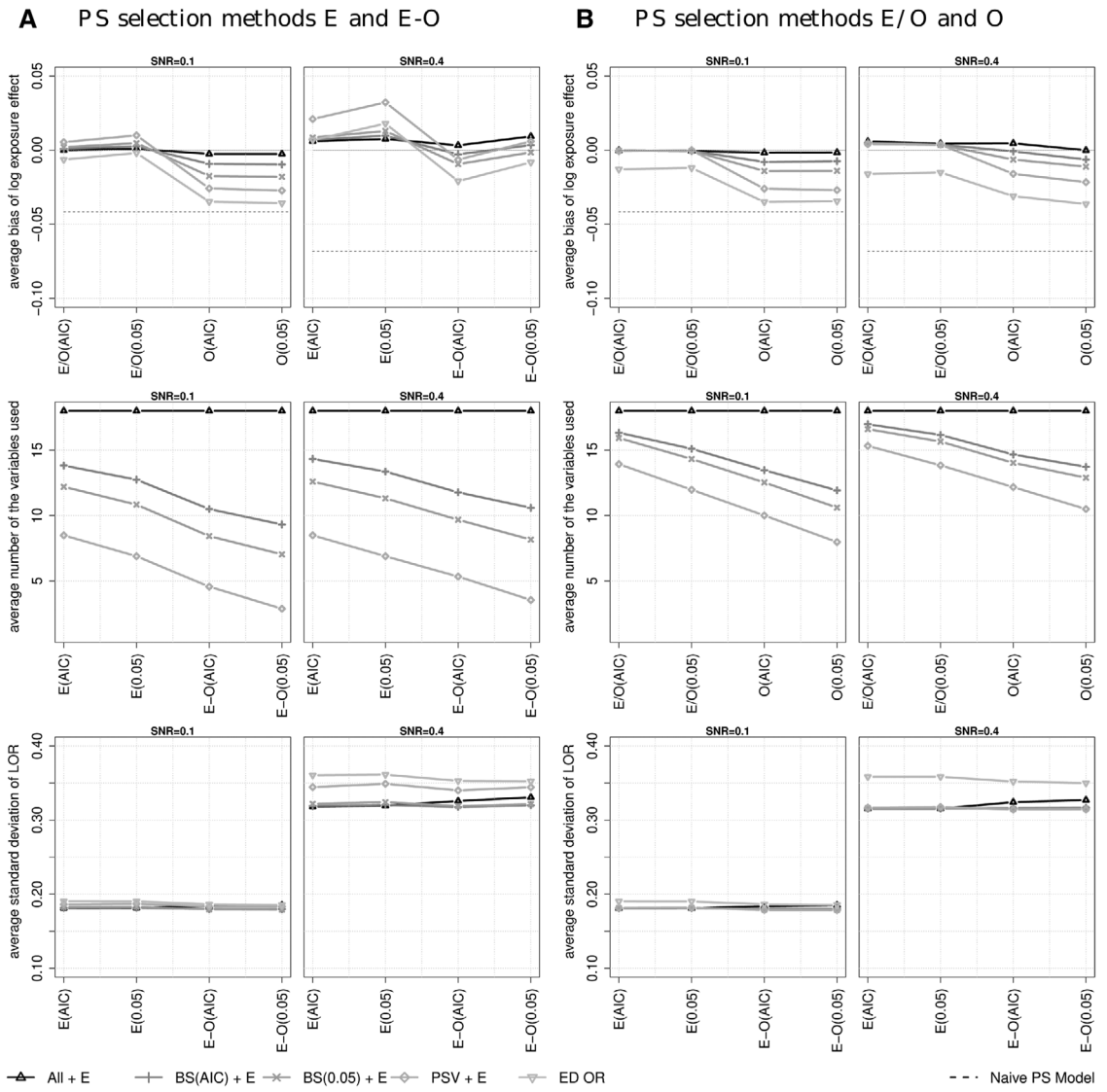

FIGURE 5 Simulation setting A-2: Covariate x_6 unavailable for selection (A) Propensity score model selected via exposure only (*E*) or both exposure and outcome (*E-O*) strategy, (B) propensity score model selected via either exposure or outcome (*E/O*) or outcome only (*O*) strategy. In the first row, the average bias of the marginal log exposure effect is given, in the second row the average number of variables used in both the propensity score model and the outcome model (Note that the result of PSV+E is equal to the result of *ED OR*), while the third row shows the average standard deviation of the marginal log exposure effect (log odds ratio, LOR). The signal-to-noise ratio (SNR) is equal to either 0.1 or 0.4. Selection models for the propensity score are displayed on the *x*-axis. Models are chosen via backward selection and with the selection threshold (AIC, *p*-value ≤ 0.05) indicated in brackets. Different symbols represent the different adjustment sets for the matched data set, whereby lighter gray indicates stricter thresholds, and black indicates no selection, meaning that we adjusted for all variables except for the unavailable one. Naive PS model denotes the combination of a propensity score model consisting of variables directly correlated with the exposure and the unadjusted post-matching outcome model

The main aim of all of the procedures used above is to correct for bias in the marginal odds ratio for exposure. The results for the setting where x_6 is unmeasured are shown in the upper panels of Figure 5. When SNR = 0.1, the bias can corrected using several approaches. Requiring only an association with the exposure to build the propensity score (E(AIC) or $E(0.05)$ leads to a good average bias reduction for all post-matching strategies. The best reduction is achieved when combining these models with a complete adjusted post-matching model ($All + E$). However, such an adjustment is often not practicable. Building the post-matching model using backward selection (BS(AIC) + E or ($BS(0.05) + E$) gives nearly the same results on average without an increase in the average standard deviation of the estimate. An even better result can be obtained when adding the variables associated with the outcome to the propensity score model (E/O(AIC) or $E/O(0.05)$. Here, the post-matching models need to at least be adjusted for the propensity score variables to fully correct the bias. The other two procedures for building the

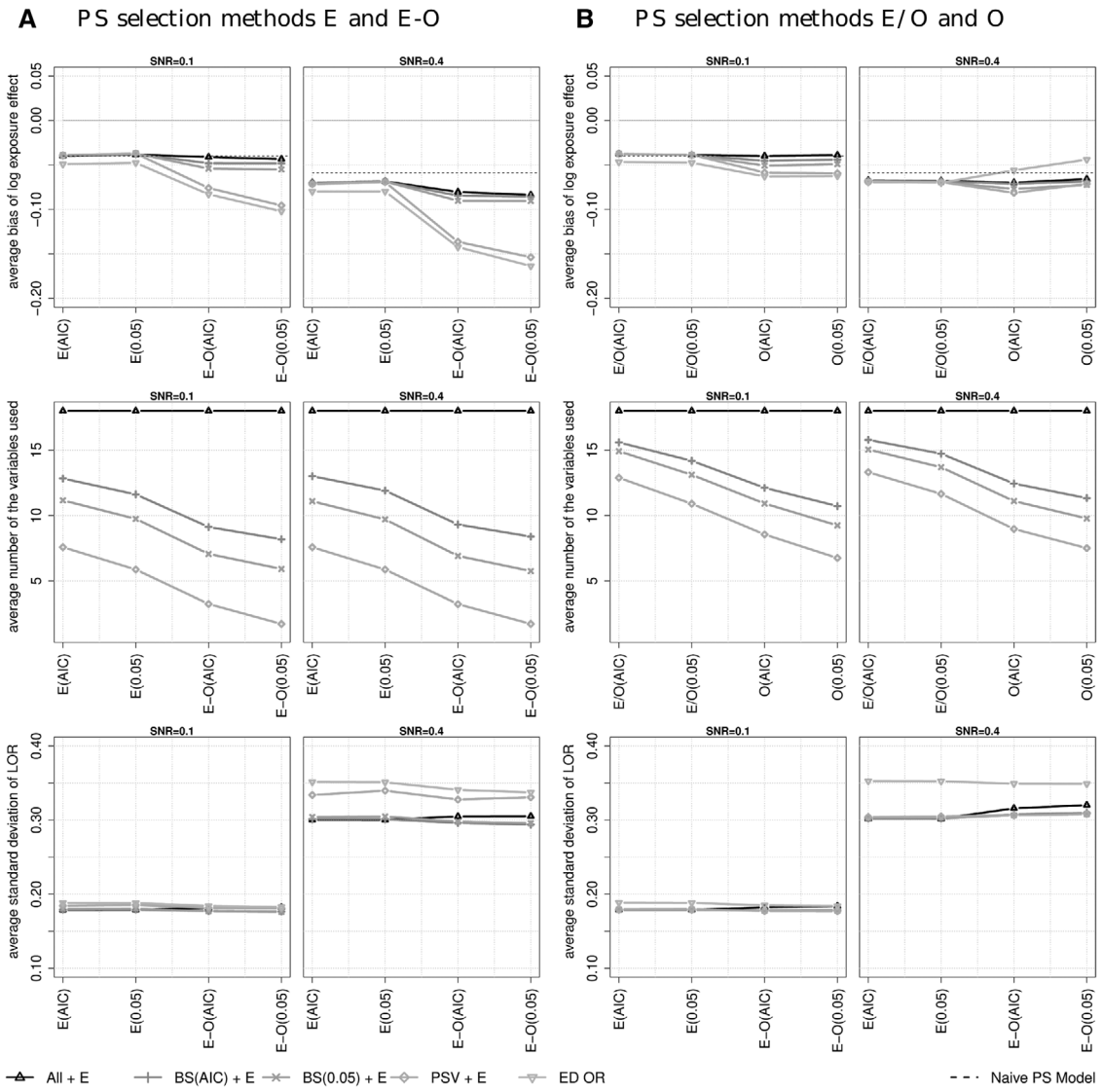

FIGURE 6 Simulation setting B-2: Covariate x_8 unavailable for selection (A) Propensity score model selected via exposure only (*E*) or both exposure and outcome (*E-O*) strategy, (B) propensity score model selected via either exposure or outcome (*E/O*) or outcome only (*O*) strategy. In the first row, the average bias of the marginal log exposure effect is given, in the second row the average number of variables used in both the propensity score model and the outcome model (Note that the result of PSV+E is equal to the result of *ED OR*), while the third row shows the average standard deviation of the marginal log exposure effect (log odds ratio, LOR). The signal-to-noise ratio (SNR) is equal to either 0.1 or 0.4. Selection models for the propensity score are given on the *x*-axis. Models are chosen via backward selection and with the selection threshold (AIC, *p*-value ≤ 0.05) indicated in brackets. Different symbols represent the different adjustment sets for the matched data set, whereby lighter gray indicates stricter thresholds, and black indicates no selection, meaning that we adjusted for all variables except for the unavailable one. Naive PS model denotes the combination of a propensity score model consisting of variables directly correlated with the exposure and the unadjusted post-matching outcome model

TABLE 5 Average number of matched individuals in the final data set

Removed	SNR	Naive *PS*	*E*(AIC)	*E*(.05)	$E - O$(AIC)	$E - O$(.05)	*E/O*(AIC)	*E/O*(.05)	*O*(AIC)	O(.05)
x_6	0.1	740	730	731	757	764	729	730	756	764
	0.4	740	730	731	759	766	729	729	757	765
x_8	0.1	740	737	738	763	770	736	737	763	770
	0.4	740	737	738	765	771	736	736	764	771

propensity score model ($E-O$(AIC), $E-O(0.05)$, O(AIC), and $O(0.05)$ lead to a larger difference between the results of the post-matching modeling approaches. Here, the more variables are used, the better the result. The bias can only be completely eliminated by using the complete post-matching outcome model, meaning that a misspecification of the post-matching outcome model can lead to substantial bias.

In the simulation setting with SNR $= 0.4$, there are four combinations where we observe a (nearly) complete bias reduction on average: (a) requiring an association with both the exposure and the outcome using the AIC for the propensity score combined with backward selection and the AIC for the outcome model ($E-O$(AIC) with BS(AIC) $+ E$); (b) using the p-value instead of the AIC ($E-O(0.05)$ with $BS(0.05)+E$); (c) only requiring an association with an outcome using AIC for the propensity score combined with backward selection using the AIC for the outcome model (O(AIC) with BS(AIC) $+ E$); and (d) using the p-value instead of the AIC for the propensity score and adjusting for all variables in the outcome ($O(0.05)$ with $ALL+E$). All of these approaches lead to a comparable standard deviation. As in the setting with SNR $= 0.1$, a misspecification of the outcome model can lead to substantial bias when requiring an association with the outcome. The greatest robustness against such a misspecification is seen when combining the variables associated with the exposure and with the outcome (E/O(AIC) or $E/O(0.05)$). However, again one needs to at least adjust for the propensity score variables. The strategy requiring only an association with the exposure (E(AIC) or $E(0.05)$ combined with backward selection for the outcome model (BS(AIC) or $BS(0.05)$ also leads to a small remaining bias while simultaneously decreasing the number of needed variables. Thus, this procedure might be the most practicable in real world data settings where the number of covariates might be high. In contrast to our expectations, all approaches with a good bias reduction led to a comparable variance of the exposure effect estimate.

To sum up, we obtain the best results across both SNR settings when we build the propensity score model by combining the variables associated with the exposure and with the outcome and also adjusting the post-matching outcome model at least for these variables. However, a high number of covariates are used in this procedure. Alternatively, an association only with the exposure selected using either the p-value or the AIC criterion for the propensity score combined with backward selection for the post-matching outcome model also gives goods results. This strategy uses fewer variables. Requiring an association with the outcome for the propensity score is not recommended as a misspecification in the post-matching outcome model can lead to substantial bias.

3.2.2 | Simulation settings B

Here, we focus on the results of the simulation setting B-2, as the results for B-1 are comparable. The latter ones can be found in Supporting Information. The covariate x_8 is not directly connected to the exposure, but affects the outcome indirectly. In the simulation setting A-2, x_8 was selected quite frequently in the propensity score model even when only requiring an association with the exposure. This was also true when x_6 was available for selection (see Appendix A in Supporting Information), underlying the importance of the covariate hypothesized in Section 2.1.

In the following we assume x_8 to be unmeasured. The propensity score strategies E and E/O lead to the same most common propensity score models as do $E-O$ and O. We can conclude that the variables deemed to be associated with the outcome are a subset of those associated with the exposure. The most common model for the different strategies comprise fewer variables than the most common model when x_6 is unavailable for the corresponding strategy. In particular, the variables x_{9a} and x_{9b} are not included although the absolute underlying nominal correlation is the same as with x_6. Thus, the strategies do not include additional variables correlated with the missing variable x_8 to approximate its influence. This inability to identify variables able to compensate for missing information is also evident from the decreased average number of variables used (Figure 6, third row). The smaller number of variables in the propensity score leads to a slightly increased number of matched individuals in the final data set (Table 5).

The removal of x_8 has a substantial effect on the average bias. The bias of the marginal log exposure effect using any of the here considered methods is at least as strong as the bias using the naive propensity score model and cannot be removed entirely. In general, there is only a small difference between the strategies and also between the two SNR settings. However, a misspecification in the post-matching outcome can again lead to substantially stronger bias when requiring an association with both the exposure and the outcome for the propensity score selection. To a lesser degree, this is also true when only requiring an association with the outcome.

As there is hardly any difference in the potential to reduce the bias, a recommendation can be based on the number of variables used and the variance in the exposure effect estimate. The following comparison focuses on the strategies with a robust bias reduction. The smallest number of variables used in total is observed when we use $E(0.05)$ for the propensity score and $BS(0.05)+E$ for the post-matching outcome model. The model building strategies hardly have any influence on the variance of the marginal exposure effect estimate.

4 | DISCUSSION

Propensity score matching is the most commonly used propensity score technique. The aim is to obtain a matched data set where the covariate distribution is the same as in the exposed individuals. Thereafter, one can interpret the marginal effect estimated using the matched cohort as the causal effect of the exposure on the exposed individuals under the SITA assumption. However, despite having been highly researched, multivariate model building is challenging and not well understood. In particular, even if one knows the actual data generation process, it still might not be clear which variables to select. In reality, this process is not fully known. Additionally, several problems such as complex correlation structures, different variable distributions, measurement errors, and unmeasured confounding are simultaneously present and potentially indistinguishable in real world data. Such a data setting, which our simulation was designed to mimic, is the main intended application of propensity score matching. Here, the true propensity score model cannot be obtained only because we added seemingly innocuous transformation techniques. To address this issue, automatic variable selection strategies can be used aiming to identify a set of confounders that satisfy the SITA assumption.

With complete data or adequate surrogate variables to compensate for the missing information, several strategies have the potential to nearly remove the bias entirely on average (simulation settings A-1 and A-2). Including variables either associated with the exposure or the outcome in the propensity score gives the best results in this setting when combined with some kind of post-matching outcome model adjustment. Nevertheless, a rather large number of variables are used, which might be problematic in situations with a high number of covariates and a limited number of observations. In general, we observe that a higher number of variables in the propensity score might lead to a lower number of matched individuals in the final data set. The number is reduced as it is not possible to identify a suitable matching partner for every exposed individual. If the sample size is low, this problem might be more pronounced and the obtained results may not be valid for the complete cohort of exposed individuals. Based on the presented simulation, one could also focus only on an association with the exposure to construct the propensity score by using backward selection procedures. In this case, we also observe good results regarding bias reduction and variance of the marginal exposure effect estimate, while simultaneously reducing the number of variables used. In contrast to other alternative strategies such as only focusing on the outcome, we observed stable results. In particular, the results were consistent across different signal-to-noise-ratios and robust against the use of the wrong post-matching outcome model selection strategy.

When an important confounder is missing without an adequate surrogate variable to approximate the missing information (simulation settings B-1 and B-2), complete removal of the bias might not possible. However, in the considered settings the propensity score modeling strategy focusing on an association with the exposure again gives good results compared to the other strategies we investigated, while keeping the number of variables used under control. Alternatively, combining this set of variables with those associated with the outcome can also be used for the propensity score model when the number of observations is sufficiently high.

In general, our results indicate that propensity score models can be performed at this step solely on the exposure while ignoring the outcome. Alternatively, one can include variables either associated with the outcome or the exposure. Compared to a strategy where only true confounders are used, that is, variables associated with both the outcome and the exposure, both of these strategies lead to less biased results in specific situations when combined with additional post-matching outcome model adjustment. This suggestion is in contrast to a previous recommendation by Patrick et al. (2011), who stated that covariates associated with the exposure, but not with the outcome, should be avoided in the propensity score model. On the other hand, it is partly in line with Austin (2007) concerning the propensity score model, but contradicts the observation of Sjölander and Greenland (2013) that no further adjustment might be needed. Our results indicate that this depends on the chosen propensity score model. For example, a propensity score model using only variables directly associated with the exposure might give biased results without post-matching adjustment. Also, post-matching adjustment can reduce the variance of the effect estimate.

To our surprise, we hardly observed any differences in the average standard deviation of the marginal exposure effect estimates between the strategies to build propensity score models. The only noteworthy differences were observed in the setting with $\mathrm{SNR} = 0.4$ when no post-matching adjustment or adjustment for propensity score variables only was performed, which resulted in a higher average variance. All other post-matching adjustment strategies resulted in comparable values, although the model size varied strongly.

The signal-to-noise rates presented correspond to correctly specified models with R^2 values of approximately 0.09 and 0.29. While these are low, they are likely to reflect the strength of relationships in real health or social science data. For instance, fully specified models from the presented case study resulted in R^2 values of 0.144 and 0.077 for survival and recurrence- or metastasis-free survival, respectively.

This simulation study has highlighted limitations in the current approach to exposure-driven propensity score matching. Current recommendations appear to not always be suitable, and consequently, more extensive simulation studies based on realistic scenarios, such as the one presented, are required. We have found that by selecting variables related to the exposure, one is more likely to minimize the impact of confounders compared to other considered modeling strategies for the propensity score without having to perform an additional post-matching adjustment. However, the latter helps to reduce the variance of the exposure effect estimate and to compensate for small misspecifications of the propensity score model. We also want to highlight that propensity score model building based solely on previous knowledge might be misleading. The real importance of some covariates might be underestimated and can only be revealed in multivariate models.

Our work has some limitations. First of all, the results cannot be generalized, but only suggest that existing recommendations might be problematic. Additionally, due to the complex simulation design, we were not able to identify a sufficient set of variables for the propensity score in the different scenarios. Nevertheless, the presented design mimics true data structures. We wanted to investigate how well the suggested methods perform in such situations to determine whether existing recommendations can be generalized. Although our results also cannot be generalized, they present a counterexample to those obtained in simpler settings. Additionally, original data often additionally contain nonlinear effects, which are also present in the original simulation design presented by Binder et al. (2011). We excluded these from the present study, but a simulation study investigating if similar results can be observed including nonlinear effects is needed. The same is true for other propensity score techniques such as weighting, or other estimating procedures such as doubly robust techniques.

To summarize, we presented different model building strategies in an exposure-driven propensity score setting in a complex, but realistic simulation design. In the simulation settings considered we observed that model building for the propensity score model solely based on the exposure might be sufficient. A strategy focusing on the association with both the exposure and the outcome might give substantially biased results when the post-matching model is misspecified. In particular, if one does not adjust for additional covariates, strong biases may be observed. These results are a counterexample to existing recommendations in the literature obtained using simple simulation designs, and show that these recommendations cannot simply be generalized to more complex settings. We conclude that more complex simulation designs than currently used are needed to give guidance for complex data structures which are expected when using observational data. Although our analysis focuses solely on one specific use of the propensity score, the results indicate that current recommendations may also be problematic when using other techniques.

ACKNOWLEDGMENTS

This work was supported by the charity German Cancer Aid (Deutsche Krebshilfe) and by the project GEnde-Sensitive Analyses of mental health trajectories and implications for prevention: A multi-cohort consortium (GESA), funded by the German Federal Ministry of Education and Research (BMBF) (FKZ 01GL1718A). It contains parts of Daniela Zöller's PhD thesis. We like to thank James Balmford for his helpful comments on the final manuscript.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID


Daniela Zöller https://orcid.org/0000-0001-9929-7403

REFERENCES

Abrahamowicz, M., Schopflocher, T., Leffondré, K., du Berger, R., & Krewski, D. (2003). Flexible modeling of exposure-response relationship between long-term average levels of particulate air pollution and mortality in the American Cancer Society study. *Journal of Toxicology and Environmental Health. Part A*, *66*, 1625–1654.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, *26*, 3078–3094.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*, 150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*, 1057–1069.

Austin, P. C. (2017). Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, *26*, 201–222.

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.

Austin, P. C., & Stafford, J. (2008). The performance of two data-generation processes for data with specified marginal treatment odds ratios. *Communications in Statistics: Simulation and Computation*, *37*, 1039–1051.

Benedetti, A., & Abrahamowicz, M. (2004). Using generalized additive models to reduce residual confounding. *Statistics in Medicine*, *23*, 3781–3801.

Beutel, M. E., Fischbeck, S., Binder, H., Blettner, M., Brähler, E., Emrich, K., … Zeissig, S. R. (2015). Depression, anxiety and quality of life in long-term survivors of malignant melanoma: A register-based cohort study. *PLoS One*, *10*, e0116440.

Binder, H., Sauerbrei, W., & Royston, P. (2011). *Multivariable model-building with continuous covariates: 1. Performance measures and simulation design*. (Report No. 105). Freiburg im Breisgau: University of Freiburg. Retrieved from http://www.fdm.uni-freiburg.de/publications-preprints/papers/pre105

Binder, H., Sauerbrei, W., & Royston, P. (2013). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: A simulation study with continuous response. *Statistics in Medicine*, *32*, 2262–2277.

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.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A*, *35*, 417–446.

Graf, E., & Schumacher, M. (2008). Comments on 'The performance of different propensity score methods for estimating marginal odds ratios' by Peter C. Austin, Statistics in Medicine 2007; 26 (16):3078–3094. *Statistics in Medicine*, *27*, 3915–3917.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 199–236.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiology and Drug Safety*, *20*, 551–559.

R Core Team (2016). R: A language and environment for statistical computing. *Retrieved from* https://www.r-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, UK: John Wiley & Sons, Ltd.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249–264.

Sauerbrei, W., & Royston, P. (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A*, *162*, 71–94.

Schmoor, C., Olschewski, M., & Schumacher, M. (1996). Randomized and non-randomized patients in clinical trials: Experiences with comprehensive cohort studies. *Statistics in Medicine*, *15*, 263–271.

Sjölander, A., & Greenland, S. (2013). Ignoring the matching variables in cohort studies - When is it valid and why? *Statistics in Medicine*, *32*, 4696–4708.

Sjölander, A., Johansson, A. L. V., Lundholm, C., Altman, D., Almqvist, C., & Pawitan, Y. (2012). Analysis of 1:1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *27*, 395–411.

Stampf, S., Graf, E., Schmoor, C., & Schumacher, M. (2010). Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. *Statistics in Medicine*, *29*, 760–769.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2005). Weaknesses of goodness-of-fit tests for evaluating propensity score models: The case of the omitted confounder. *Pharmacoepidemiology and Drug Safety*, *14*, 227–238.

Wynant, W., & Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in Medicine*, *33*, 3318–3337.

Wyss, R., Girman, C. J., LoCasale, R. J., Alan Brookhart, M., & Stürmer, T. (2013). Variable selection for propensity score models when estimating treatment effects on multiple outcomes: A simulation study. *Pharmacoepidemiology and Drug Safety*, *22*, 77–85.

Zöller, D., Wockner, L., & Binder, H. (2020). Modified ART study - Simulation design for an artifical but realistic human study dataset. *Retrieved from* https://doi.org/10.5281/zenodo.3678736

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zöller D, Wockner LF, Binder H. Automatic variable selection for exposure-driven propensity score matching with unmeasured confounders. *Biometrical Journal*. 2020;62:868–884. https://doi.org/10.1002/bimj.201800190