

The impact of unmeasured confounding in observational studies: a plasmode simulation study of targeted maximum likelihood estimation

Lateef Amusa^{1,2*}, Temesgen Zewotir¹, Delia North¹

¹ Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

² Department of Statistics, University of Ilorin, Ilorin, Nigeria

* Corresponding author: Lateef Amusa, amusasuxes@gmail.com, +27659904686, Department of Statistics, University of Kwazulu-Natal, Durban, South Africa

Abstract

Unmeasured confounding can cause considerable problems in observational studies and may threaten the validity of the estimates of causal treatment effects. There has been discussion on the amount of bias in treatment effect estimates that can occur due to unmeasured confounding. We investigate the robustness of a relatively new causal inference technique, targeted maximum likelihood estimation (TMLE), in terms of its robustness to the impact of unmeasured confounders. We benchmark TMLE's performance with the inverse probability of treatment weighting (IPW) method. We utilize a plasmode-like simulation based on variables and parameters from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). We evaluated the accuracy and precision of the estimated treatment effects. Though TMLE performed better in most of the scenarios considered, our simulation study results suggest that both methods performed reasonably well in estimating the marginal odds ratio, in the presence of unmeasured confounding. Nonetheless, the only remedy to unobserved confounding is controlling for as many as available covariates in an observational study, because not even TMLE can provide safeguard against bias from unmeasured confounders.

Keywords: TMLE; IPW; odds ratio; unobserved confounding; plasmode simulation.

Background

Observational studies of health care interventions and treatments are critical for assessing treatment effects and making patient-centred treatment decisions. Apart from the infeasibility of conducting randomized experiments due to the cost and ethical reasons, observational studies, unlike randomized trials, answer questions of safety and comparative effectiveness that are of direct interest to health care providers and patients faced with a treatment decision [1]. In some randomized trials, patients are exposed to treatments without fully understanding their benefits and risks [2].

The increasingly available observational data sources, including large health care databases, provide an opportunity to efficiently evaluate new treatments and compare with existing alternatives. Like every observational study, these evaluations require confounding control or adjustment. Confounders are referred to as variables or factors associated with both treatment assignment and outcome. While randomized experiments, via the randomization mechanism, balance both measured and unmeasured confounders between treatment groups, observational studies usually suffer from confounding effects. When substantial imbalances exist between the treated and control groups in an observational study, biased treatment effect estimates are produced.

Various techniques, including multivariable regression adjustment, propensity score matching, stratification and weighting, have been commonly used to account for observed or measured confounders [3]. However, these techniques are not entirely immune to the effects of unmeasured confounding, which is not uncommon in clinical and medical researches; hence it is still a vexing problem. For instance, a large cohort study found evidence of the effectiveness of ascorbic acid (vitamin C) in reducing all-cause mortality by 52% (relative risk: 0.48; 95% CI: 0.33 - 0.70) [4],

after adjusting for age, sex, body mass index, smoking status, systolic blood pressure, serum cholesterol, diabetes, and vitamin C supplement use. This effect was, however, not found in randomized trials (relative risk: 1.00; 95% CI: 0.94 - 1.06, for vitamin C supplementation versus placebo) [5]; hence it was disputed whether the disparity of the results was due to the observational study being biased due to unmeasured confounding by, say, socioeconomic status or dietary habits [6-8].

Though several strategies, including sensitivity analysis, have been proposed and adopted to address the potential impact of unobserved confounders on causal inference estimation, they are not model estimation methods and do not have inbuilt facilities of determining the robustness to the impact of uncontrolled confounders. Sensitivity analysis is a systematic assessment of the changes in model estimation when the potential unobserved confounders are present versus when they are absent. This technique has been widely and increasingly utilized [9-11].

Despite extensive studies of the performance of the methods for confounding control, the robustness of these methods to uncontrolled confounders has received almost no attention. Much recently, targeted maximum likelihood estimation (TMLE) - a doubly-robust semiparametric method, has gained the attention of applied researchers [12-15]. TMLE estimates exposure effects or associations without relying on model specification. It combines semiparametric estimation, using machine learning algorithms, with an additional estimation process to optimize a parameter of interest (e.g. risk difference, risk ratio, and odds ratio) [16, 17]. We sought to examine the performance of TMLE in terms of its robustness to the impact of unmeasured confounders on estimated treatment or exposure effects.

Using the inverse probability of treatment weighting (IPW) method, the most commonly used weighting method by applied researchers and practitioners, especially in the medical and health

sciences, as a benchmark, we evaluated TMLE in a plasmode-based simulation study. Our plasmode simulation was inspired by a real empirical cohort study that is typical of health care interventions and treatments. The plasmode approach preserves the number and type of observed covariates in the original study, as well as the correlation structure among covariates, and attempts to simulate exposure and outcome to reflect the observed associations of these variables with those from the real data [18].

Methods

Assume W to be a vector of covariates, and T, Y denote the exposure (or treatment) indicator and observed outcome, respectively. Let $Y(T = t) = Y(t)$ indicate an individual's potential outcome under treatment $T=t$. A valid causal inference using IPW or TMLE requires some necessary assumptions. First, the time ordering assumption is required: the covariates W precede treatment T , T precedes Y in time, and Y depends on T and W , while T depends only on W . Secondly is the consistency assumption, which requires that an individual's potential outcome under the actually received treatment equals the observed outcome; $Y(T = t) = (Y | T = t)$. The third is the positivity assumption, which requires that (within strata of W), every individual had a non-zero probability of receiving one of the two treatment conditions, i.e., $0 < P(T = 1|W) < 1$. Finally, the assumption of no unmeasured confounding (also known as the conditional exchangeability assumption) is of ultimate interest in this study. This assumption states that the potential outcomes $Y(t)$ are independent of the observed treatment T given measured covariates; $Y(T = t) \perp T / W$. It implies that (within the strata of W), the outcome under the potential treatment $T = 1$, i.e, $P(Y(1) = 1 | T = 1, W)$ equals the outcome under treatment condition $T = 0$, i.e, $P(Y(1) = 1 | T = 0, W)$. In other words, all confounders have been measured.

Targeted maximum likelihood estimation

We briefly describe the implementation of TMLE with binary outcome and exposure or treatment for the estimation of the marginal odds ratio. Two probability models are initially specified and estimated based on the observed data on individuals $i = 1, 2, \dots, n$. First, the initial conditional probability of outcome Y , given the treatment and covariates, $Q_0(T, W) = E_0(Y | T, W)$ is estimated. Second, the conditional distribution of the treatment, given covariates W , $g_0 = P(T | W)$, is estimated.

The estimate $Q_n^0(T_i, W_i)$ and the predictions $Q_n^0(1, W_i)$ and $Q_n^0(0, W_i)$ can be estimated with the standard logistic regression; machine learning tools like the Super Learner may also be used. Super Learner is an ensemble learner of a pre-specified library of algorithms with parameters. It uses cross-validation to adaptively create an optimally weighted combination of estimates from candidate algorithms [19].

The estimates $Q_n^0(1, W_i)$, and $Q_n^0(0, W_i)$ are then plugged-in into the substitution estimator of the parameter of interest, log odds ratio, to obtain an untargeted estimate:

$$\Psi_{MLE,n} = \text{logit} \{Q_n^0(1, W_i)\} - \text{logit} \{Q_n^0(0, W_i)\},$$

Where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$.

Initial estimates of $Q_0(T, W)$ are then updated along a path of some fluctuation parameters, incorporating additional information from the propensity score function to reduce residual confounding in $Q_0(T, W)$. This updating involves two steps: Firstly, g_n , is used in a clever covariate $H_n^*(T, W)$ to define a parametric working model to fluctuate $Q_0(T, W)$.

$$H_n^*(T, W) = \left(\frac{I(T=1)}{g_n(1|W)} - \frac{I(T=0)}{g_n(0|W)} \right)$$

For each individual with $T_i = 1$ and $T_i = 0$, the clever covariates are calculated as $H_n^* (1, W_i) = \frac{1}{g_n(1|W_i)}$ and $H_n^* (0, W_i) = \frac{-1}{g_n(0|W_i)}$, respectively. In addition to adding the columns $H_n^* (1, W_i)$ and $H_n^* (0, W_i)$, these values are then combined to form a column $H_n^* (T_i, W_i)$ in the data matrix.

In the final step, the fluctuation parameter ϵ_n is estimated by fitting an intercept-free logistic regression of Y on $H_n^* (T, W)$ with the logit of $Q_n^0 (T, W)$ being an offset (fixed quantity), where ϵ_n is the resulting coefficient of the clever covariate $H_n^* (T, W)$. Accordingly, the estimate Q_n^0 is updated into a new estimate Q_n^1 of $Q_1 (T, W)$:

$$\text{Logit } Q_n^1 (T, W) = \text{Logit } Q_n^0 (T, W) + \epsilon_n H_n^* (T, W).$$

Consequently, the predictions from the updated model are used to estimate the probabilities of the counterfactual outcomes:

$$\text{Logit } Q_n^1 (1, W) = \text{Logit } Q_n^0 (1, W) + \epsilon_n H_n^* (1, W),$$

and

$$\text{Logit } Q_n^1 (0, W) = \text{Logit } Q_n^0 (0, W) + \epsilon_n H_n^* (0, W),$$

The updated estimates $Q_n^1 (1, W)$ and $Q_n^1 (0, W)$ are empirically averaged and used to compute the targeted estimator the marginal odds ratio:

$$\Psi_{TMLE,n} = \frac{\left[\frac{1}{n} \sum_{i=1}^n Q_n^1 (1, W_i) \right] \left[1 - \frac{1}{n} \sum_{i=1}^n Q_n^1 (0, W_i) \right]}{\left[\frac{1}{n} \sum_{i=1}^n Q_n^1 (0, W_i) \right] \left[1 - \frac{1}{n} \sum_{i=1}^n Q_n^1 (1, W_i) \right]}$$

The standard error can be estimated with the efficient influence function [20], and the Wald-type confidence intervals can be calculated correspondingly.

Inverse probability of treatment weighting

Marginal treatment effects can be estimated in the presence of confounding with the IPW estimator. IPW reduces confounding by correcting each i th individual's contribution by weight equal to the inverse of the estimated probability of exposure or treatment receipt [21]. This

probability is usually estimated by logistic regression. By applying IP weighting, a pseudo-population, where the distribution of covariates is comparable across treatment groups, is created [22]. Therefore, by contrasting the marginal outcome between the two treatment conditions in the created pseudo-population, via a weighted unadjusted logistic regression for the outcome, an unbiased estimate of the marginal treatment effect is produced. We used a robust sandwich-type variance estimator [22] to estimate the standard error of the IPW estimator.

Simulation study

Description of simulated datasets

We adopted a plasmode-like simulation design to emulate real-world datasets with complex correlation and effect structures. This approach allows us to specify different prediction models for the exposure and outcome variables based on the covariance matrix and original estimates of the regression coefficients. We based our simulations on a previously published cohort study on the effectiveness of right heart catheterization (RHC) for critically ill patients [23, 24]. The study, known as the SUPPORT study, was a five-centre study of decision making and outcomes in hospitalized, severely sick, hospitalized adult patients. Amongst other things, we chose this cohort due to its relatively high number of confounder variables, which can thus provide a challenging dataset for selecting potential confounders.

In brief, the dataset, comprises information on 5735 patients, 2184 (38%) of them being treated with RHC ($T_i=1$) within 24 hours of admission in an intensive care unit, and 3551 (72%) assigned to the control group ($T_i=0$); the outcome of interest was mortality at 30 days of admission. A total of 1918 patients (33.4%) died within 30 days. Full details of this data, including the variable description and its summary statistics, have been published elsewhere [24-26].

The full set of potential confounders W (49 pre-specified variables) from the original data, were retained, and the estimated coefficients (66 estimated parameters) from the data were used as parameters of the simulation. Both the treatment and outcome variables were generated based on logistic regression models: 1) for the treatment variable T as a function of the covariates W

$$\text{logit } P(T = 1 | W) = \alpha_0 + \sum_k^{66} \alpha_k W_k,$$

2) for the observed outcome as a function of the treatment variable and the covariates W .

$$\text{logit } P(Y = 1 | W) = \beta_0 + \beta_T T + \sum_k^{66} \beta_k W_k.$$

The parameter values for intercepts α_0 and β_0 were unaltered to preserve the treatment and outcome prevalences in the real data.

To study the impact of unmeasured confounding, we assumed the model that adjusts for the full covariate set, to contain the true treatment effect. We then partition the covariate space into two subsets of potential confounders that were used to assess the robustness of TMLE to the impact of not adjusting for the full covariate set. In the first set (W_1), also known as the simple set, we adjusted for predefined important confounders, including demographic characteristics. These covariates include sex, age, race, primary disease category, type of insurance, and APACHE score. In the second set, known as the moderate set (W_2), we included variables from the simple set (W_1) and combined literature knowledge with variables associated with RHC use from the original analysis [24]. These covariates include cancer status, support model estimate of surviving two months, comorbidities, mean blood pressure, white blood cell count, albumin, sodium, pH levels, do-not-resuscitate DNR status on day 1, heart rate, respiratory rate, and temperature. In each of the two covariate sets, other measured confounders in the full set were intentionally omitted.

Apart from the degree of unmeasured confounding (the two covariate sets size), other factors we varied were sample size ($n = 1000, 5000$) and confounding weight (1, 1.15, 1.3, 1.45). By confounding weight, we mean the multipliers that were multiplied to the estimated covariate coefficients to increase the amount of confounding in the simulated data. In total, we have 2 sample sizes x 4 confounding weights x 2 covariate sets = 16 different simulation scenarios. We then simulated two potential outcomes: one assuming that the patient received the RHC treatment $Y(1)$ and one assuming that the patient did not receive the RHC treatment $Y(0)$. For each of the 16 scenarios, we generated 1000 simulated datasets by repeating this process 1000 times, and we averaged two sets of potential outcomes and calculated the true marginal odds ratio.

Analyses of simulated datasets

After simulation of the datasets, we investigated the impact of unmeasured confounding on TMLE, relative to IPW, in terms of accuracy and precision of treatment effect estimates. For each of the simulation scenarios, we applied each of the TMLE and IPW methods of treatment effect estimation. For TMLE, due to computational simplicity, we used generalized linear models (GLMs) to estimate both the PS and outcome models. For both methods, the covariates were included as main terms in both the PS and outcome model. Interactions between covariates were not considered.

For each method, we estimated the marginal odds ratio and computed the bias, mean squared error (MSE), model-based standard errors, and 95% confidence interval (CI) coverage of the estimated treatment effects. CI coverage is the proportion of times the estimated confidence intervals contain the specified parameter value. TMLE was implemented using the *tmle* package [27] in R version 4.0.0.

Results

We present the simulation results based on the scenarios and performance metrics explained in the earlier section. We focus on the performance of the TMLE method, using IPW as a threshold for evaluating the results. For both estimation approaches, except for the 95% CI coverage, there were no substantial differences in the bias, MSE, and standard error of estimated treatment effects, given the small and moderate covariate set adjustments. In all considered scenarios, it was observed that the MSE increased as the confounding weight increases, irrespective of the sample size. The same pattern was observed for the standard error and 95% CI coverage. For the bias, an opposite pattern was observed. The absolute bias was consistently decreasing with increasing confounding weights. However, this pattern was clear and distinct for the large sample size ($n = 5000$) only.

In Figure 1, the results of the absolute bias of estimated treatment effects are presented for the different scenarios explained in the earlier section. The bias is generally low in the moderate-sized sample size (values ranged from 0.013 – 0.240) and is generally high given high sample size (values ranged from 0.119 – 0.158). For moderate sample size ($n = 1000$), TMLE produced slightly lower bias only when the confounding strength was higher than 1.15. However, no notable difference was observed given a large sample size ($n = 1000$).

TMLE maintained lower MSE between the two methods, for moderate sample ($n = 1000$), with the difference in MSE being increasingly larger as the confounding weight increased (Figure 2). However, indistinguishable estimates were observed for large sample size ($n = 5000$).

In Figure 3, the results of the model-based standard error of estimated treatment effects are presented for the different scenarios explained in the earlier section. The TMLE standard errors were incredibly smaller in all the scenarios (TMLE range: 0.005 – 0.116, IPW range: 0.072 –

0.337). Much smaller standard errors were observed in the large sample size ($n = 1000$) scenarios. In terms of the 95% CI coverage, there was no clear pattern of superiority between the two methods; both techniques generally achieved reasonably high coverages overall (Figure 4). For moderate sample size ($n = 1000$), both methods produced 95% CI above the nominal coverage rate, while for large sample size ($n = 5000$), higher coverage rates were achieved at confounding weights higher than 1.15.

Figure 1: Bias of estimated treatment effects. Top panels: $n = 1000$, A) small covariate set scenarios. B) moderate covariate set scenarios. Bottom panels: $n = 5000$, C) small covariate set scenarios. D) moderate covariate set scenarios

Figure 2: Mean squared error of estimated treatment effects. Top panels: $n = 1000$, A) small covariate set scenarios. B) moderate covariate set scenarios. Bottom panels: $n = 5000$, C) small covariate set scenarios. D) moderate covariate set scenarios

Discussion

Unmeasured confounding may threaten the validity of the estimates of causal treatment effects. An excellent causal inference in observational studies is as good as the controlled, observed confounders. Accordingly, an excellent observational study strives to control for a rich set of covariates - as many as available, so as to minimize the risk of unmeasured confounding. In this study, we investigate a relatively new method of estimating causal treatment effects, TMLE, in terms of its robustness to the impact of unmeasured confounding. Using a plasmode-like simulation study, we compared its performance with IPW, a relatively popular and commonly used method in applied studies. While focusing on TMLE, we summarize our findings and place them in the context of existing literature, where necessary.

Figure 3: Model-based standard error of estimated treatment effects. Top panels: $n = 1000$, A) small covariate set scenarios. B) moderate covariate set scenarios. Bottom panels: $n = 5000$, C) small covariate set scenarios. D) moderate covariate set scenarios

Though TMLE performed better in most cases, our simulation study results suggest that both methods performed reasonably well in estimating the marginal odds ratio, in the presence of unmeasured confounding. The outperformance of TMLE was most striking for the model-based standard errors. In most scenarios, there were no substantial differences in the performance of estimated treatment effects, given the small and moderate covariate set adjustments. In other words, increasing (or decreasing) the number of unobserved confounders did not substantially change the estimates. Except for the standard error, in all the other performance metrics, including bias, mean squared error, and 95% confidence interval coverage, the performance of both methods were indistinguishable for large sample size ($n = 5000$).

Figure 4: 95% confidence interval coverage of estimated treatment effects. Top panels: $n = 1000$, A) small covariate set scenarios. B) moderate covariate set scenarios. Bottom panels: $n = 5000$, C) small covariate set scenarios. D) moderate covariate set scenarios

We used parametric models for the initial treatment and outcome models estimation due to computational simplicity. However, to take optimal advantage of the properties of TMLE, data-adaptive methods, such as Super Learner [19], are recommended when implementing TMLE. The potential integration of data-adaptive methods while retaining valid inference is a unique strength of TMLE.

Our study is not devoid of limitations. Although we covered a number of scenarios, our simulated data were based on one real cohort, exemplifying only one type of complex data structure from

the data. Even a Monte Carlo simulation study can never capture the whole space of possible data scenarios as they occur in reality, the generalizability of results is a potential limitation in our plasmode simulation study. Assignment of a broad range of covariate distributions or parameter values is not usually possible; the covariates parameter values are instead adopted directly from a real-world setting and fixed for simulation. Therefore, our results cannot necessarily be generalized to settings that have not been evaluated; hence we cannot draw definitive conclusions on the comparative performance of the TMLE and IPW estimators. However, this specific data setting is typical of medical and health studies and has already provided useful insights and evaluation regarding the performance of TMLE in the presence of (multiple) unmeasured confounders. Additionally, we did not explore other outcome types, including continuous and time-to-event outcomes, in the context of this study. We look forward to filling the gaps left by these limitations.

In conclusion, unmeasured confounding is still a vexing problem in observational studies. Not even TMLE can provide safeguard against bias from unmeasured confounders. Therefore, more efforts should be devoted to collecting and controlling for as many as available covariates in an observational study. If data on any confounder is available, it should be controlled for in the estimation of effects. For instance, in a study of the relationship between plasma ascorbic acid level and mortality, it was noted that, despite being recorded, information on the social class and physical activity of the participants was not used in the analysis [4]. Lawlor and colleagues [7] showed that if these variables were indeed confounders of the ascorbic acid and mortality association, even a moderate effect of any of them would have resulted in considerable residual confounding in the reported estimates. We further support the recommendation that sensitivity

analysis of (multiple) unmeasured confounders should routinely perform to guide the discussion on the possible direction, magnitude, and its impact [9, 28-31].

List of abbreviations

HIV: TMLE: targeted maximum likelihood estimation; IPW: inverse probability of treatment weighting

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The dataset used in this study is available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable No funding was received for this research

Authors' contributions

LA designed the study and analyzed the data. TZ and DN critically reviewed the manuscript and gave constructive comments which improved the manuscript. All authors have read and approved the manuscript.

Acknowledgement

Not Applicable

References

1. Goldberg NH, Schneeweiss S, Kowal MK, Gagne JJ. Availability of comparative efficacy data at the time of drug approval in the United States. *JAMA*. 2011;305(17):1786-9.
2. Halpern SD, Ubel PA, Berlin JA, Townsend RR, Asch DA. Physicians' preferences for active-controlled versus placebo-controlled trials of new antihypertensive drugs. *Journal of general internal medicine*. 2002;17(9):689-95.
3. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*. 2010;25(1):1.
4. Khaw K-T, Bingham S, Welch A, Luben R, Wareham N, Oakes S, et al. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *The lancet*. 2001;357(9257):657-63.
5. Group HPSC. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20 536 high-risk individuals: a randomised placebo-controlled trial. *The Lancet*. 2002;360(9326):23-33.
6. Concato J, Horwitz RI. Observational versus randomised trial evidence. *The Lancet*. 2004;364(9436):756-7.
7. Lawlor DA, Smith GD, Bruckdorfer KR, Kundu D, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *The Lancet*. 2004;363(9422):1724-7.
8. Lawlor D, Ebrahim S, Kundu D, Bruckdorfer K, Whincup P, Smith GD. Vitamin C is not associated with coronary heart disease risk once life course socioeconomic position is taken into account: prospective findings from the British Women's Heart and Health Study. *Heart*. 2005;91(8):1086-7.
9. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;948-63.
10. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *Statistical models in epidemiology, the environment, and clinical trials*: Springer; 2000. p. 1-94.
11. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and drug safety*. 2006;15(5):291-303.
12. Bodnar LM, Cartus AR, Kirkpatrick SI, Himes KP, Kennedy EH, Simhan HN, et al. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. *The American Journal of Clinical Nutrition*. 2020;111(6):1235-43.
13. Herrera R, Berger U, Von Ehrenstein OS, Díaz I, Huber S, Moraga Muñoz D, et al. Estimating the Causal Impact of Proximity to Gold and Copper Mines on Respiratory Diseases in Chilean Children: An Application of Targeted Maximum Likelihood Estimation. *International journal of environmental research and public health*. 2018;15(1):39.
14. Lim S, Tellez M, Ismail AI. Estimating a Dynamic Effect of Soda Intake on Pediatric Dental Caries Using Targeted Maximum Likelihood Estimation Method. *Caries research*. 2019;53(4):486-94.
15. Rodríguez-Molina D, Barth S, Herrera R, Rossmann C, Radon K, Karnowski V. An educational intervention to improve knowledge about prevention against occupational asthma and

allergies using targeted maximum likelihood estimation. *International archives of occupational and environmental health*. 2019;92(5):629-38.

16. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*. 2017;185(1):65-73.

17. Van Der Laan MJ, Rubin D. Targeted maximum likelihood learning. *The international journal of biostatistics*. 2006;2(1).

18. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis*. 2014;72:219-26.

19. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Statistical applications in genetics and molecular biology*. 2007;6(1).

20. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data: Springer Science & Business Media; 2011.

21. Rosenbaum PR. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.

22. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. LWW; 2000.

23. Murphy DJ, Cluff LE. The SUPPORT study. *Journal of Clinical Epidemiology*. 1990;43:V-X.

24. Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Jama*. 1996;276(11):889-97.

25. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-99.

26. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*. 2001;2(3-4):259-78.

27. Gruber S, Van der Laan MJ. *tmle: An R package for targeted maximum likelihood estimation*. 2011.

28. Greenland S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2005;168(2):267-306.

29. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*. 1959;22(1):173-203.

30. Greenland S. Basic methods for sensitivity analysis of biases. *International journal of epidemiology*. 1996;25(6):1107-16.

31. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003;451-8.