BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Identification of confounders and estimating the causal effect of place of birth on age-specific childhood vaccination

Ashagrie Sharew Iyassu[1,5*], Haile Mekonnen Fenta[1,2,3], Zelalem G. Dessie[1,4] and Temesgen T. Zewotir[4]

Abstract

Background In causal analyses, some third factor may distort the relationship between the exposure and the outcome variables under study, which gives spurious results. In this case, treatment groups and control groups that receive and do not receive the exposure are different from one another in some other essential variables, called confounders.

Method Place of birth was used as exposure variable and age-specific childhood vaccination status was used as outcome variables. Three approaches of confounder selection techniques such as all pre-treatment covariates, outcome cause covariates, and common cause covariates were proposed. Multiple logistic regression was used to estimate the propensity score for inverse probability treatment weighting (IPTW) confounder adjustment techniques. The proportional odds model was used to estimate the causal effect of place of birth on age-specific childhood vaccination. To validate the result obtained from observed data, we used a plasmode simulation of resampling 1000 samples from actual data 500 times.

Result Outcome cause and common cause confounder identification techniques gave comparable results in terms of treatment effect in the plasmode data. However, outcome causes that contain common causes and predictors of the outcome confounder identification gave relatively better treatment effect results. The treatment effect result in the IPTW confounder adjustment method was better than that of the regression adjustment method. The effect of place of birth on log odds of cumulative probability of age-specific childhood vaccination was 0.36 with odds ratio of 1.43 for higher level vaccination status.

Conclusion It is essential to use plasmode simulation data to validate the reproducibility of the proposed methods on the observed data. It is important to use outcome-cause covariates to adjust their confounding effect on the outcome. Using inverse probability treatment weighting gives unbiased treatment effect results as compared to the regression method of confounder adjustment. Institutional delivery increases the likelihood of childhood vaccination at the recommended schedule.

Keywords Confounder, Causal inference, Plasmode simulation, Place of birth, Vaccination

*Correspondence:
Ashagrie Sharew Iyassu
statashe@gmail.com
[1]College of Science, Bahir Dar University, Bahir Dar, Ethiopia
[2]Center for Environmental and Respiratory Health Research, Population Health, University of Oulu, Oulu, Finland
[3]Biocenter, University of Oulu, Oulu, Finland
[4]School of Mathematics, Statistics & Computer Science, University of KwaZulu Natal, Durban, South Africa
[5]Debremarkos University, Debre Markos, Ethiopia

Background

In a randomized control trial, "ideal methodology for causal inference" [1], random allocation of exposure controls both known and unknown pre-exposure variations of subjects that may influence the outcome. However, due to ethical and practical issues, randomized control trials are used in limited ranges [1]. An alternative non-randomized research design, called an observational study, is used to estimate the causal effect of an exposure on the outcome [2, 3]. Observational studies may lack internal validity since associations between exposure and outcome may be biased due to other factors that act as a confounder or selection bias [4].

Confounding is an issue in nearly all observational studies that focus on causality. In causal analyses, the relationship between the exposure and the outcome variables under study can be altered by some other third factor. In this case, treatment groups and control groups that receive and do not receive the exposure are different from one another in some other essential variable that is also associated with the outcome [5]. Thus, a variable that alters the relationship between exposure, which is the potential cause of the outcome, and the outcome, is considered a confounder.

According to [6, 7], three characteristics should be satisfied for a variable to be a confounder. It should be associated with exposure and outcome; it must be distributed unequally between treatment and control groups; and it should not be in the causal pathway between exposure and outcome. Temporally, confounders come prior to exposure. A covariate that comes after exposure has taken place is not a confounder since it is unable to retroactively modify the exposure [2].

In observational study, confounders have to be identified and controlled their effect while estimating the association between exposure and outcome. Controlling confounders helps to get unbiased estimates of the exposure-outcome relationship [8]. How to identify confounders and how to deal with them is one of the challenges in observational study. There is no common consensus criteria to identify which covariates are confounders and which are not [9]. A common approach is to control as many pre-exposure covariates as possible [2]. Studies modified this approach by controlling all covariates significantly associated (p-value less than 0.05) with the outcome of interest [10, 11] mentioned in [2]. Others also stated that control confounders give a predetermined magnitude of change (10% or 15%) while estimating the relationship between exposure and outcome [10, 12]. Directed acyclic graph (DAG) is also a method of identifying confounders to be controlled in the association of exposure with the outcome. It aims to identify a minimally satisfactory set of well-measured confounders that satisfy the definition of confounders to control [2]. Confounders can be selected by using statistical grounds such that 10% or more of a covariate changes an estimate, forward and backward variable selection criteria, and machine learning methods [5].

Despite controversies about which method is better, change in estimate and significance testing methods are widely used to confounder identification even if significance testing is acceptable for a P-value of 0.2 or less [9]. On the other hand, the use of change in estimate is questionable due to its low ability to enhance the precision of treatment effect estimate [13]. Thus, the two significance approaches such as common cause, outcome cause, and all pre-treatment covariate approaches were used. To the level of our knowledge, confounders that influence the causal effect of place of birth on age-specific childhood complete vaccination and the level of its effect after controlling confounders has not been documented.

Hence, this study was conducted to identify confounders that result spurious associations between place of birth (exposure) and childhood vaccination (outcome). In addition, the causal effect of an exposure on the outcome was estimated by controlling confounders using regression and inverse probability treatment weighting (IPTW). Plasmode simulation [14] was done to see the reproducibility of the identification method and effect of exposure on the outcome with such real data.

Method

Study setting and description of data

The data used in this study was obtained from the Ethiopian Mini Demographic and Health Survey (EMDHS) 2019. Specifically, data from birth records that contain all records of women aged 15–49 with the most recent birth prior to five years of the survey were used. The data was collected from March 21, 2019 to June 28, 2019 from nationally representative samples using two- stage stratified sampling to provide estimates at the national and regional levels as well as for urban and rural areas. In the survey, 5,753 women with live births were interviewed [15]. However, only children who were alive at the time of the survey were considered for this study.

In this study, the exposure variable was mother's place of birth and the outcome variable was age-specific childhood vaccination status. Age-specific childhood vaccination is when mothers or caregivers vaccinate their child as per the recommendation of the Ethiopian expanded program on immunization [16]. When a child received a particular vaccination at the right age, a score of 1 was given, otherwise, 0 was given. If a child receives all vaccines timely, it is labeled as fully vaccinated. If one or more vaccines were missed at each age, it is labeled as partially vaccinated and labeled as no vaccination when a child took no vaccination at each age. A total of 5,150 surveyed children were involved in this study.

Pre-exposure covariates such as mothers' characteristics, child characteristics, and household characteristics were taken as covariates that possibly confound the association between exposures and outcome. The possible relationship between covariates, exposure and outcome is presented in the DAG given in Fig. 1. DAG is a visual representation of the assumed causal mechanisms and can help to identify covariates to adjust for confounding and control confounding bias (Fig. 1). The arrows in the DAG show the direction of causal relationships.

Notation and assumption

Let A be the binary exposure with values of 0 and 1, $\boldsymbol{X}$ be a possible confounders and $Y(a)$ is the potential outcome associated with treatment, $A = a$. The following assumptions should be kept for the unbiased causal effect of the treatment on the outcome.

Unconfounded assumption: This is also called ignorability assumption where the treatment assignment is independent of the potential outcome given the set of measured covariate: $Y(a) \perp A/\boldsymbol{X}$ [17]. When propensity score is used as covariate adjustment rather than conditioning on them, $\boldsymbol{X}$ can be replaced by propensity score denoted by $e(x)$.

Positivity assumption: $0 < P(A = a/X) < 1, \forall a \epsilon 0, 1$, this is an overlap or common support assumption that requires every study participant to have a chance to be in any of the treatment conditions. The probability of treatment assignment for any participant is neither zero nor one under treatment conditions [18].

Stable unit treatment value assumption (SUTVA): the potential outcomes of i^{th} subject are not influenced by the potential outcome of j^{th} subject for $i \neq j$, and each unit receives the same version of the treatment [19].

Potential outcome framework and treatment effect

For causal inference under the potential outcome framework [20], the potential outcome is the possible outcome under different treatment conditions. For binary treatment, such as place of delivery, $A = \{0,1\}$, a subject has two potential outcome, $Y(1)$ for $A = 1$ (when a subject received treatment) and $Y(0)$ when $A = 0$ (when a subject does not receive treatment). The outcome is observed only under one, and not under both treatment conditions. If a subject receives treatment level $A = 1$, then $Y(1)$ is observed but $Y(0)$ is unobserved; if a subject receives treatment level $A = 0$, then $Y(0)$ is observed but $Y(1)$ is unobserved. Under the potential outcome framework, the observed outcome is written as:

$$Y = (1 - T) * Y(0) + T * Y(1)$$

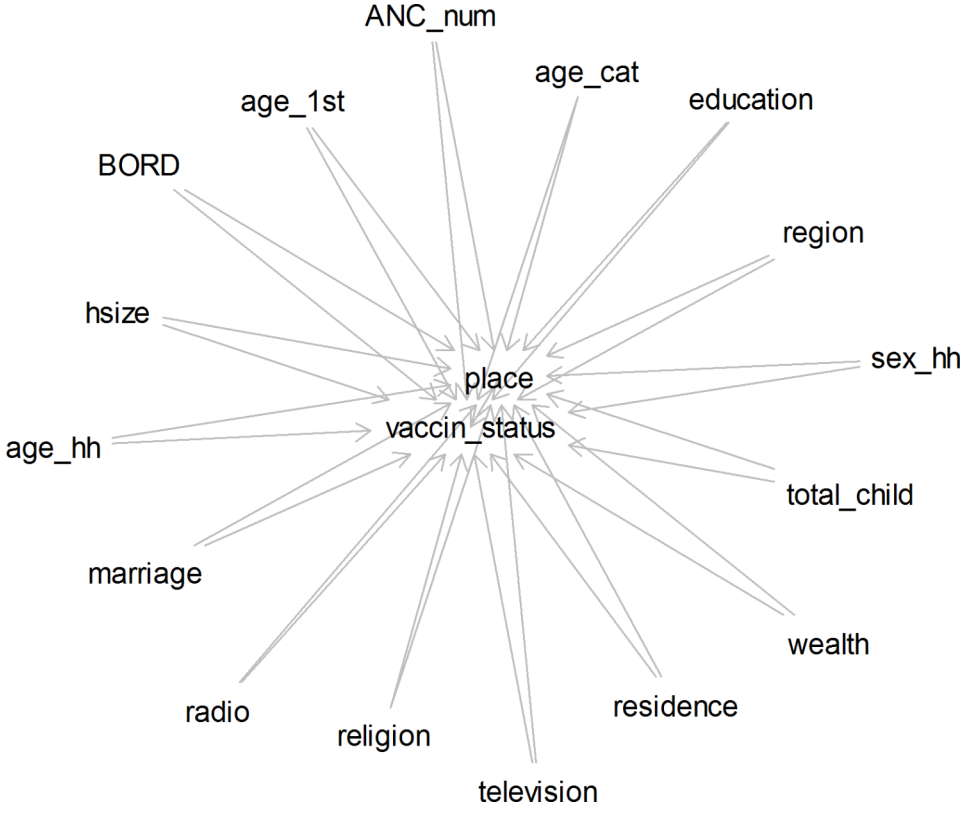

Fig. 1 DAG for covariates, exposure and outcome. The description of covariates, exposure and outcome is given in the supplementary material

.

The treatment effect at individual level is $Y_i(1) - Y_i(0)$.

Because it is impossible to calculate change of individual-level treatment effect, we need to estimate average or population-level treatment effects change under the assumption of unconfoundness or exchangeability. Thus, the average treatment effect between treated and untreated subjects is denoted and given by: $\Delta\, ATE = E(Y(1)) - E(Y(0))$.

In the presence of measured baseline covariates that could result spurious relationship between treatment (treatment and exposure alternatively used), two approaches were used to control confounding effects: the first approach uses ordinal regression model conditional on treatment and confounders, and the second approach uses inverse probability treatment weighting which is marginal structural modeling.

Let Y_j, $j = 0,1,2$ be the status of age-specific childhood vaccination, a be place of delivery (0 for home delivery and 1 for institutional delivery), then the proportional odds model is given by [21, 22]:

$$\begin{aligned}\text{For regression approach}: \ & \mathrm{g}\left(\mathrm{E}\left(\mathrm{p}(\mathrm{Y} < \mathrm{j}/\mathrm{a}, \mathbf{X})\right)\right) \\ & = \alpha_j - (\beta_1 a + B^T{}_2 \boldsymbol{X}).\end{aligned} \tag{1}$$

where $\mathrm{g}(.)$ is the link function of the proportional odds model.

$$\mathrm{ATE} = \mathrm{g}\left(\mathrm{E}\left(\mathrm{p}(\mathrm{Y} < \mathrm{j}/1, \mathbf{X})\right)\right) - \mathrm{g}\left(\mathrm{E}\left(\mathrm{p}(\mathrm{Y} < \mathrm{j}/0, \mathbf{X})\right)\right)$$

$$\begin{aligned}\text{For marginal structural modeling}: \ & \mathrm{g}\left(\mathrm{E}\left(\mathrm{p}(\mathrm{Y} < \mathrm{j}/\mathrm{a})\right)\right) \\ & = \alpha_j - \beta_1 a\end{aligned} \tag{2}$$

In this case, each value of the subject is weighted by IPTW using WeightIt R package [23] and covariate balancing was checked using cobalt R package [24].

Identification of confounders

Different criteria were used to select cofounders. The first was pre-exposure criteria [25]. In this criterion, one can consider and control all covariates that come prior to exposure under study. The second criterion was significance testing criteria (ST) with cutoff P-values fixed at less than or equal to 0.2. In this criterion, two approaches were used: common cause [26] that states a covariate is a confounder when it causes the exposure and outcome; outcome cause that states one can consider a covariate to be a confounder when it is a cause of the outcome. The outcome causes contained covariates associated with both exposure and outcome and covariates associated the outcome but unrelated to the exposure [17]. Using the notation used by Shortreed & Ertefaie (2017), let $\mathrm{X_l}$ be a set of all pre-treatment covariates, X_p be a set of common cause covariates, and X_q be a set of outcome cause covariates. These three different covariates were included in Eq. (1) while using the regression method to adjust confounders and estimate treatment effect. While using Eq. (2) to estimate the treatment effect, we used propensity score to control confounders estimated after fitting the generalized linear model given as follows:

$$g\left(E(P(A = 1/\boldsymbol{X}; \widehat{\theta}))\right) = \widehat{\theta}_0 + \widehat{\boldsymbol{\theta}}^T \boldsymbol{X} \tag{3}$$

$g(.)$ is the link function for binary response such as logit link functions, and $\mathbf{X}$ is one of $\mathrm{X_l}$, X_p and X_q, and $\widehat{\boldsymbol{\theta}}$ is a vector of estimated coefficients.

Plasmode simulation

The most common data generation approaches are parametric and plasmode simulation [27]. In parametric simulation, covariates, exposure and outcome data are generated from known or predefined stochastic distributions such that the generated data resembles the realistic or representative. Parameters of interest are derived from the real data, literature, or set by the user [28, 29]. Parametric simulation is mainly used for model development and to compare the performance of different models.

Plasmode simulation starts resampling of exposure and covariates from the original data. Then the outcome data is generated from resampled exposure and covariates. The parameters or effect sizes in the simulation process are estimated from the original data by modeling the relationship of outcome with exposure and covariates. Sometimes the parameters can be defined by the user. Plasmode simulation is viewed as semi-parametric simulation because exposure and covariates are generated naturally from unknown parametric distribution whereas; the outcome is simulated from known distributions and resampled exposure and covariate [14, 27]. The approach depends on resampling from the practical exposure and covariates data without modification in all simulated datasets to preserve the relations among these variables and complex data structure. It has also the advantage of a user-specified exposure effect. Moreover, the simulated datasets are used to compare variable selection strategies for confounder adjustment via the propensity score [13].

The plasmode data were simulated [27], with the following procedures:

I. Exposure and covariates structures are generated by resampling from an original data.
II. Generating outcome that incorporates.

   II.1 Determining outcome generating model: linear regression with normal distribution was used and categorized into ordered categories.

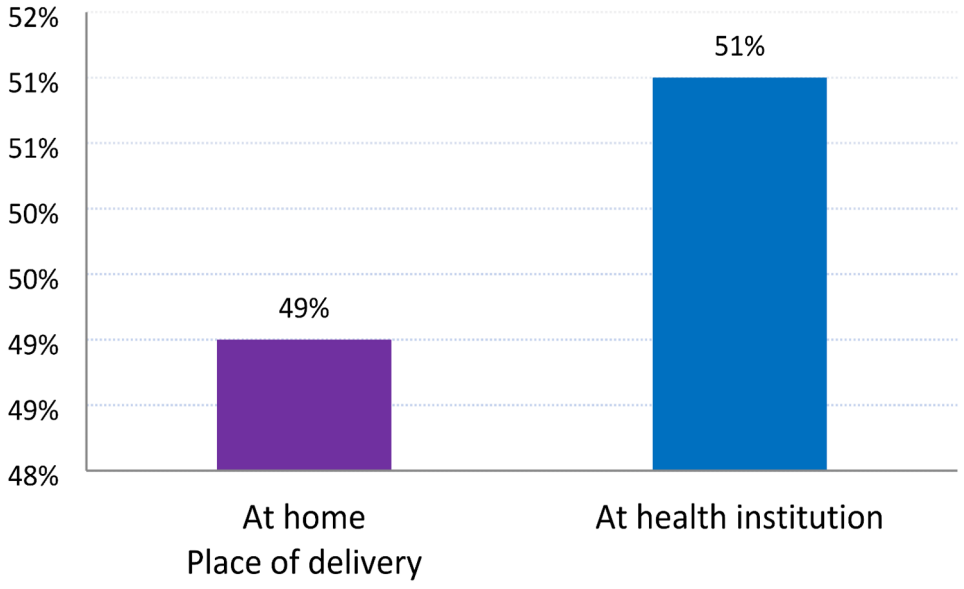

Fig. 2 Distribution of place of delivery

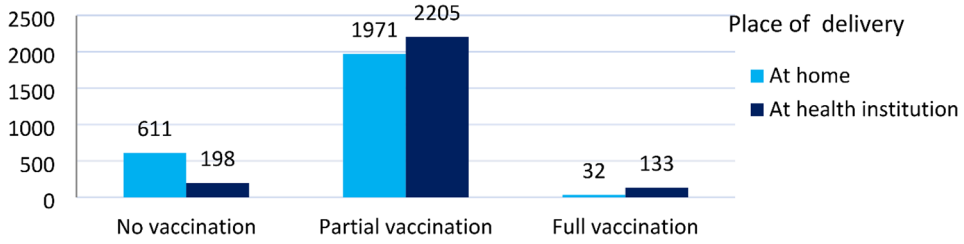

Fig. 3 Distribution of vaccination status along with place of delivery

II.2. Determining exposure and covariate effect by estimation from original data and individual specification.
II.3. Generating outcome data using chosen outcome model, effects, and resampled covariates and exposure.

The challenges of plasmode simulation are specifying the number of plasmode data sets (N), resampling techniques (Resampling without replacement, m out of n, and resampling with replacement, n out of n), and resampling size (m). So far there is no defined rule to choose the sampling techniques, to determine simulation and resample sizes [27]. In this study, the size of the simulated data set or the number of repetitions (N) was specified to be 500 as used in [14, 27, 30, 31]. Sampling without replacement technique was implemented. The m out of n resampling with replacement also called subsampling requires fewer assumptions and prevents bootstrap failure [27, 32]. In addition, the size of the resample (m) was 1000.

Comparison of methods

To compare the performance of methods from simulated data, we used different multi-dimensional performance measurement metrics illustrated by [29]. These include bias, mean square error, empirical standard error, average model standard error, and bias eliminated coverage. For each performance measurement metrics estimate, Monte Carlo standard error was also estimated. The estimates of each performance measures from simulation study was generated using simsum R function from rsimsum package [33, 34] The estimand of the study was treatment effect on the outcome.

Result

Distribution of place of birth and its association with the outcome

The percentage distribution of place of birth is visualized in Fig. 2. It is shown that, out of 5150 mothers, 2614 (51%) of them delivered at home and 2536 (49.24%) of them delivered at health institutions.

On the other hand, the frequency distribution of association of place of delivery with the outcome (age-specific childhood vaccination status) is visualized in Fig. 3. Among 4176 age-specific partially vaccinated children, 52.8% (2205) were delivered at health facilities and the rest 47.2% (1971) of them were delivered at home.

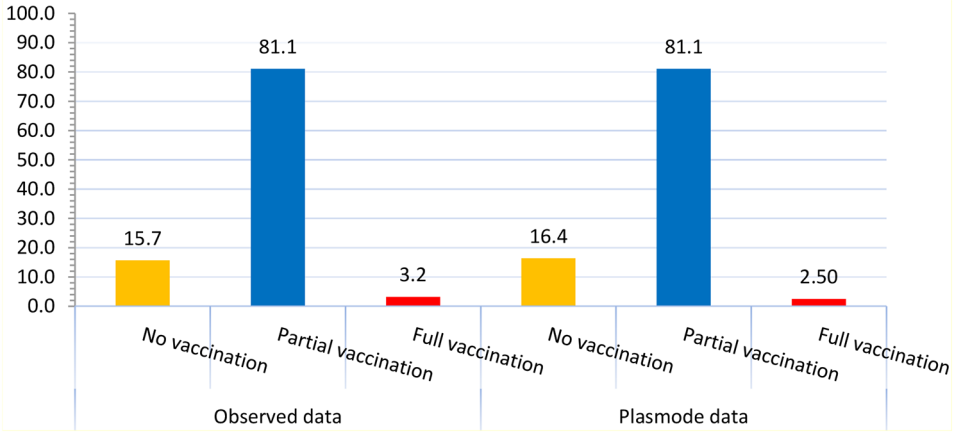

Fig. 4 Distribution of plasmode and observed data

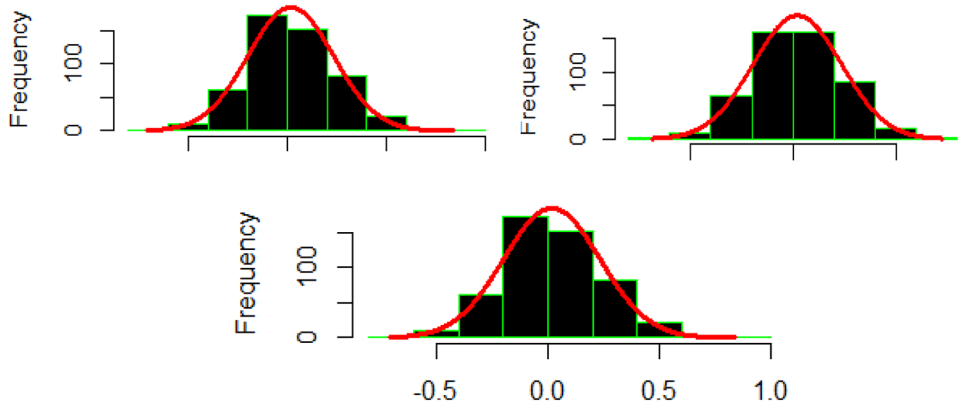

Fig. 5 Histogram of treatment effect estimate from plasmode simulation for the three confounder identification techniques

Similarly, from no age-specific vaccinated children, 809, only smaller (24.5%, 198) of them were delivered at health facilities, and larger (75.5%, 611) of them were delivered at home. From on time fully vaccinated children, 165, large proportions (80.6%, 133) of them were delivered at health institutions (Fig. 3).

Result of plasmode simulation

To make sure plasmode simulation generates realistic data that resemble the observed data, we compared the distribution of the simulated and observed data. We simulated 500 artificial data sets for the outcome each containing 1000 mothers with the proportion of childhood vaccination status matched with the observed data. In Fig. 4, we presented the bar chart for the distribution of each category of the outcome. It shows the proportions of each category are similar between simulated and observed data sets.

Evaluation of confounder selection strategy

We used simulated data for the covariate selection strategy to be included in propensity score function or to control with the regression method. Figure 5 is the histogram of treatment effect estimator distribution from simulation study. The distribution of the estimator for each confounder selection techniques indicates, the estimator is nearly symmetrical. Thus, we are able to estimate other method performance metrics assuming the estimator normally distributed. On the other hand, any of the estimators from simulation study was equal to the true treatment effect which was 0.5.

Figures 6 and 7 also show the result of treatment effect for each simulation step of proposed covariate selection approaches. Figure 6 presents the treatment effect before adjusting confounders, outcome cause (**out.cause**), common cause (**com.cause**), and all pre-treatment covariates (**All.covariates**) adjusted by regression. The distribution of crude effects is far from the three approaches. However, the distribution of treatment effect is similar for the three approaches. Figure 7 also shows similar result after adjusting with propensity score-based IPTW.

Table 1 demonstrates the estimate of method performance metrics for simulation study using inverse treatment probability weighting from 500 simulated data sets for each confounder selection approaches. In the generation of artificial outcomes from real covariates, we fixed

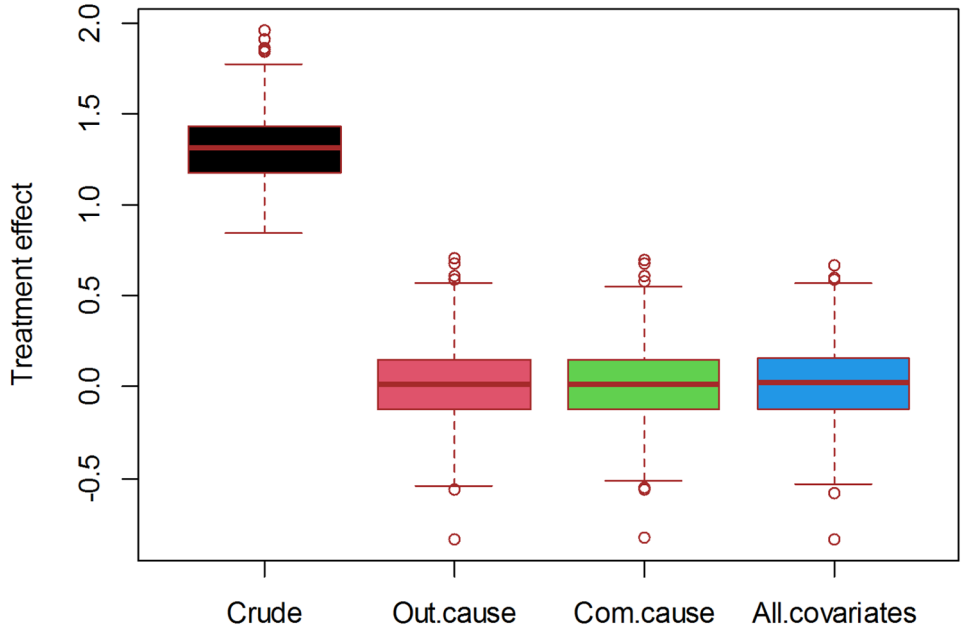

Fig. 6 Treatment effect for plasmode data with regression covariate adjustment

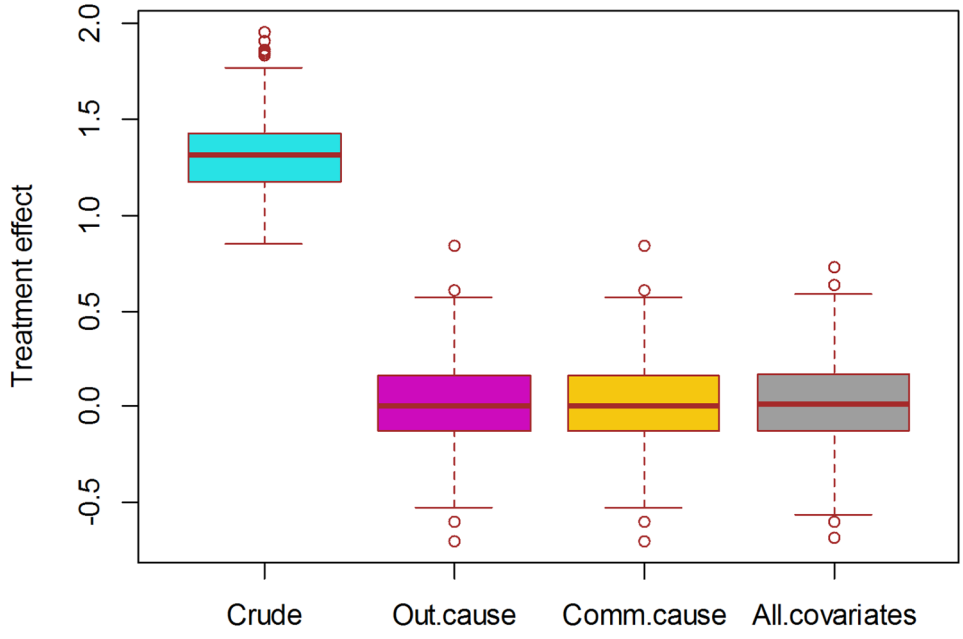

Fig. 7 Treatment effect from plasmode simulated data based on IPTW covariate adjustment

the true treatment effect at 0.5 for all approaches and confounder adjustment methods. The average treatment effect after adjusting confounders was almost similar for all approaches with negligible difference between outcome and common cause. The true treatment effect was highly reduced when taking confounders into account and adjusting with IPTW.

The absolute value of bias for all-pretreatment covariates was higher than outcome and common cause covariates. Its uncertainty (Monte Carlo simulation study) was also higher than the other two approaches. The absolute difference of bias between outcome and common cause covariates was 0.001 with approximately equal uncertainty values. The bias eliminated coverage of outcome and common cause covariates was 88.8% and that of all pre-treatment covariates was 89.2% with equal Monte Carlo standard error for all approaches. The empirical standard error which is the square root of the variance of an estimator and measure of the efficiency of an estimator was equal in outcome cause and all-pre-treatment

Table 1 Estimates of performance measures for confounder identification approaches using plasmode simulation, values in parentheses are Monte Carlo standard errors when confounders are adjusted with IPTW

Performance measure	Outcome cause	Common cause	All pre-treatment
Average treatment effect estimate	0.019	0.018	0.022
Bias	-0.481 (0.01)	-0.482 (0.01)	-0.978 (0.02)
Coverage	0.236(0.0190)	0.250 (0.019)	0.22 (0.002)
Bias eliminated coverage	0.888 (0.014)	0.888 (0.014)	0.892 (0.014)
Empirical standard error	0.216(0.007)	0.218(0.007)	0.216 (0.007)
MSE	0.2784 (0.0097)	0.2794 (0.0098)	1.0028 (0.0190)
Model based standard error	0.167 (0.0003)	0.167 (0.0003)	0.168 (0.0003)
Relative % error in standard error	-22.83 (2.45)	-23.48(2.43)	-22.20 (2.47)
Power of 5% test	0.114 (0.014)	0.116 (0.014)	0.108 (0.014)

Table 2 Estimates of performance measures for confounder identification approaches using plasmode simulation, values in parentheses are Monte Carlo standard errors when covariates are adjusted with regression

Performance measure	Outcome cause	Common cause	All pre-treatment
Average treatment effect estimate	0.020	0.021	0.021
Bias	-0.780 (0.0094)	-0.791 (0.0094)	-0.790 (0.0097)
Coverage	0.040 (0.0088)	0.040 (0.0088)	0.050 (0.0097)
Bias eliminated coverage	0.960 (0.0088)	0.958 (0.0090)	0.954 (0.0094)
Empirical standard error	0.211 (0.0067)	0.211 (0.0067)	0.217 (0.0069)
MSE	0.652(0.0150)	0.653 (0.0150)	0.654 (0.0155)
Model based standard error	0.209(0.0003)	0.208 (0.0003)	0.213 (0.0003)
Relative % error in standard error	-0.847 (3.1422)	-1.399 (3.1246)	-1.72 (3.1147)
Power of 5% test	0.042 (0.0090)	0.046 (0.0094)	0.052 (0.0099)

covariates. The empirical standard error of common cause was little higher than other approaches. On the other hand, the model based standard error which is the average of the standard error of estimators from 500 simulation data was equal in outcome and common cause confounder selection approaches. However, its value was little higher in all pre-treatment confounder selection approach than outcome and common cause confounder selection approaches. The magnitude of relative percent error in standard error in common cause was higher than the other two approaches. The difference of power of test at 5% between outcome and common cause confounder selection approaches was very small (difference=0.002). Considering all method performance measures and taking the aggregates, outcome cause confounder identification approach is better than common cause and all pre-treatment covariates approach.

Table 2 also shows the estimate of average treatment effect and performance measurement metrics for confounder identification approaches when confounders were adjusted using regression. The bias of treatment effect for outcome cause covariates was little smaller than the other two approaches (Table 2). The bias eliminated coverage of outcome cause covariates somehow more than the other two approaches. Similarly, the estimator in outcome cause covariates was more efficient than all pre-treatment covariates with smaller values of empirical standard error and model based standard error.

The magnitude of relative percent error in standard error, and the power of test at 5% test of outcome cause covariates was smaller than the other two approaches (Table 2). The result when confounders were adjusted using regression method implies that outcome cause covariate performed better than the common cause and all pre-treatment covariates.

On the other hand, the bias and mean square error of adjusting confounders with IPTW was smaller than adjusting confounders with regression method.

Table 3 Effect of place of birth on age-specific childhood vaccination when confounders are adjusted by regression and IPTW

Confounder adjustment method	Confounder selection approach	Average treatment effect	Standard error of treatment effect	95% CI	
Regression	Unadjusted treatment effect	1.3	0.081	1.16	1.5
	Outcome cause	0.53	0.099	0.34	0.73
	Common cause	0.54	0.099	0.34	0.73
	All pre-treatment covariates	0.51	0.101	0.32	0.71
IPTW	Outcome cause	0.36	0.072	0.23	0.50
	Common cause	0.37	0.073	0.23	0.52
	All pre-treatment covariates	0.46	0.075	0.31	0.60

Result from observed data

Table 3 contains the effect of treatment (place of birth) on outcome (age-specific childhood vaccination) along with their standard errors for unadjusted, outcome cause, common cause, and all pre-treatment covariates adjusted by regression and IPTW. Before using IPTW for marginal structural model, we checked covariate balance and positivity of propensity score for the three confounder selection approaches. Figure S1-S3 in the supplementary

material shows covariate balance using absolute standardized mean difference. Figure S4 also shows positivity of propensity scores that range from 0.021 to 0.999.

Unadjusted treatment effect on the log odds of the cumulative probability of the outcome was 1.3. When adjusted for outcome cause, common cause, and all pre-treatment covariates using the cumulative link model, the effect dropped down to 0.53, 0.54, and 0.51 respectively. When using IPTW, the effect dropped down to 0.36, 0.37 and 0.46 respectively. To compare confounder selection approaches, almost all brought the same reduction of crude effect in the regression technique although it looks like all pre-treatment gave the largest reduction with higher standard error. When using IPTW, the outcome cause approach brought the largest reduction with relatively the smallest standard error. On the other hand, IPTW reduced the unadjusted treatment effect with a smaller standard error and narrow confidence interval as compared to the regression method. This shows that the IPTW confounder adjustment method is preferable to the regression method. Based on outcome cause confounder selection approach, number of antenatal care services, age of household head, age at first birth, household size, total number of children ever born, birth order number, region, place of residence, religion, mother's education status, ownership of television and radio, and household wealth status were the confounders for the causal effect of place of delivery on age-specific childhood vaccination.

Effect of place of birth on age-specific childhood vaccination

After identifying and adjusting confounders, the next step was estimating the treatment effect on the outcome. Choosing outcome cause approach for confounder selection, and using the IPTW confounder adjustment method (Tables 1 and 2), the log odds of institutional delivery (coded as 1) versus home delivery was given as follows:

$$logit(P\,(y \leq\ j/T) = \alpha_{\,j} - 0.36T,\ \ T = 0{,}1$$

Then the odds ratio of the event, $y \leq\ j\ is,\ OR\ = exp^{-0.36} = 0.70$ and that of $y > j\ is,\ OR = exp^{0.36} = 1.43$. This implies that institutional delivery decreases the lower level of vaccination status (no vaccination and partial vaccination) with 30%. On the other hand, institutional delivery increases the likelihood of higher level of vaccination status (partial from no vaccination and full vaccination from partial vaccination) with 43%.

Discussion and conclusion

Discussion

This study was done to identify confounders and estimate the causal effect of place of birth on age-specific childhood vaccination. The treatment/exposure (alternatively used) variable was place of delivery and the outcome variable was age-specific childhood vaccination categorized as no, partial/incomplete, and full/complete vaccination. In the process of estimating exposure's causal effect on the outcome, we have to be curious about extraneous/cofounding covariates that could alter the effect of the exposure on the outcome. Most studies so far focused on confounder identification and treatment effect estimation when the outcome is continuous or binary. However, to the best of our knowledge, no literature that dealt with when the outcome is ordinal, especially the causal effect of place of birth on age-specific childhood vaccination.

The study used regression and propensity score-based inverse probability treatment weighting to correct confounding effects. A common approach is to control as many covariates as possible that are observed before the treatment [2]. However, including all pre-treatment covariates in any confounder controlling methods such as regression introduces bias. Adding more covariates to the model causes over fitting and unstable coefficients due to multicollinearity [35]. A model is best when it contains the smallest number of covariates that explain the greatest amount of variance [9]. To prove this assertion, we proposed three approaches of confounder selection to control their effect. These are all pre-treatment covariates, the common cause of treatment and outcome covariates (real confounders) and outcome cause covariates. As described in [17, 36], the outcome cause covariates included real cofounders and predictors of outcome variable.

To show the reproducibility of our confounder selection to be included in the regression model and propensity score estimation, we used plasmode simulation as used in [13, 27]. A key advantage of statistical plasmode simulations is their ability to maintain the intricate structure of real-world data by resampling covariate information from an actual dataset. For these simulations to be effective, it is essential to have a suitable representative dataset, which forms the foundation for the entire plasmode simulation study [27]. We resampled 1000 data sets without replacement 500 times out of 5150 total data sets. A latent continuous artificial outcome was simulated from a known covariate structure and manually added truth treatment effect (to be constant across confounder selection approaches). The effect of each covariate was determined from the actual relationship of observed data. The performance of each confounder selection approaches was compared using multi-dimensional performance

metrics as described by Morris et al. [29] using IPTW and regression based confounder adjusting.

The result from plasmode data set showed that there were no notable differences across confounder selection approaches but it looks outcome cause approach gives better results. On the other hand, common cause covariates are the subset of outcome cause covariates. Hence, we used outcome cause covariates to estimate the causal effect of place of delivery on timely childhood vaccination. Even though the regression and IPTW gave comparable results, it seems IPTW can give better results for confounder adjustment. However, in the plasmode data, there was a huge difference between the true treatment effect and the treatment effect after covariate adjustment which is too far from the treatment effect from observed data. One of the challenges to generating artificial ordinal outcomes is that there is no direct method to simulate it. We rather simulate continuous latent outcome and then cut in to ordered categories based on the quintiles of observed data. In this process, there might be loss of information, and the treatment effect in observed and simulated data could deviate.

The result from the observed data set, outcome cause gave relatively better result as compared to other approaches. This is consistent with the findings in the literature [14, 17, 36]. The limitation of confounder identification using a regression method is it does not show us how treatment and control groups are balanced in terms of baseline covariates. On the other hand, the propensity score helps us to evaluate how confounders are balanced between treatment and control groups. Hence, the result of this study showed that IPTW gave better result for treatment effect with better confounders balancing.

Finally, after identifying confounders using outcome cause approach and adjusting them using IPTW, the causal effect of place of birth on age-specific childhood vaccination was estimated. The result shows institutional delivery enhances the timely vaccination of children.

Conclusion

In causal inference, the identification of confounders is an important step before attempting to identify exposure's effect on outcome. In this study, we considered the place of delivery as binary exposure that causes age-specific childhood vaccination which is the ordinal outcome. Adding all pre-treatment covariates in the regression model or propensity score function could result in spurious treatment effect on the outcome. Hence, it is important to identify confounders and adjust them with propensity scores, or other matching methods. In the study, outcome cause covariates (common causes and predictors of the outcome) results relatively better performance than other proposed methods in terms of treatment effect. Because if we include covariates related to the treatment and the outcome, there can be other covariate(s) that are not related to the treatment but have a mixing effect with the treatment on the outcome. Propensity score inverse probability treatment weighting helps to see the balanced confounders between treatment and control groups. It also gave better treatment effect as compared to regression adjustment of covariates as demonstrated in this study. Institutional delivery enhances the likelihood of a newborn baby to get vaccine as per the standard schedule.

Abbreviations

ATE	Average treatment effect
DAG	Directed acyclic graph
EMDHS	Ethiopian mini demographic and health survey
IPTW	Inverse probability treatment weighting
SUTVA	Stable unit treatment value assumption

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02827-2.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

A.S analyze data and write manuscriptH.M , Z.D and T.Z revised and edited the manuscript.

Funding

Authors did not receive any fund for this study.

Data availability

The data was obtained from the Ethiopian mini demographic and health survey (EMDHS) collected from March 21, 2019, to June 28, 2019. It was accessed from DHS program website: https://dhsprogram.com/data/available-datasets.cfm.

Declarations

Research ethics and consent to participant

We used secondary data which was collected by Ethiopian Public Health Institute in collaboration with the Central Statistical Agency and the Federal Ministry of Health. Hence, the study does not need ethical clearance and patient consent with two reasons. We, the researchers, did not involve in data collection instead the data was collected by authorized governmental institutions keeping all ethics and patient consent. Second, the result and report of this study does not harm study participants or no personal issue that we disclose.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 April 2024 / Accepted: 16 December 2024
Published online: 27 December 2024

References

1. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med. 2018;210:2–21.
2. Ranapurwala SI. Identifying and addressing confounding bias in violence prevention research. Curr Epidemiol Rep. 2019;6:200–7.
3. Tennant PW, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021;50(2):620–32.
4. Lash TJ. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
5. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 2019;34:211–9.
6. Rothman KJ, Greenland S, Lash TL. Modern epidemiology, Vol. 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia; 2008.
7. Porta M. A dictionary of epidemiology. Oxford University Press; 2014.
8. Lee PH, Burstyn I. Identification of confounder in epidemiologic data contaminated by measurement error in covariates. BMC Med Res Methodol. 2016;16:1–18.
9. Tong S, Lu Y. Identification of confounders in the assessment of the relationship between lead exposure and child development. Ann Epidemiol. 2001;11(1):38–45.
10. Wiebe DJ. Homicide and suicide risks associated with firearms in the home: a national case-control study. Ann Emerg Med. 2003;41(6):771–82.
11. Culyba AJ, et al. Association of future orientation with violence perpetration among male youths in low-resource neighborhoods. JAMA Pediatr. 2018;172(9):877–9.
12. Branas CC, et al. Investigating the link between gun possession and gun assault. Am J Public Health. 2009;99(11):2034–40.
13. Talbot D, et al. The change in estimate method for selecting confounders: a simulation study. Stat Methods Med Res. 2021;30(9):2032–44.
14. Franklin JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219–26.
15. EPHI I. Ethiopian Public Health Institute (EPHI)[Ethiopia] and ICF. Ethiopia Mini Demographic and Health Survey 2019: Key Indicators; 2019.
16. Health FMo. Ethiopia national expanded programme on immunization. BMJ Publishing Group FMOE Addis Ababa; 2015.
17. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. Biometrics. 2017;73(4):1111–22.
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
19. Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press; 2015.
20. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc. 2005;100(469):322–31.
21. Christensen RHB. Cumulative link models for ordinal regression with the R package ordinal. Submitted J Stat Softw. 2018;35.
22. Smith TJ, Walker DA, McKenna CM. An exploration of link functions used in ordinal regression. J Mod Appl Stat Methods. 2020;18(1):20.
23. Greifer N. Using weightit to estimate balancing weights.
24. Greifer N. Covariate balance tables and plots: a guide to the cobalt package, vol. 10. Accessed March, 2020. p. 2020.
25. Rubin DB. For objective causal inference, design trumps analysis; 2008.
26. Glymour MM, Weuve J, Chen JT. Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias. Neuropsychol Rev. 2008;18:194–213.
27. Schreck N, et al. Statistical plasmode simulations–Potentials, challenges and recommendations. Stat Med. 2024;43(9):1804–25.
28. Burton A, et al. The design of simulation studies in medical statistics. Stat Med. 2006;25(24):4279–92.
29. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102.
30. Atiquzzaman M, et al. Using external data to incorporate unmeasured confounders: a plasmode simulation study comparing alternative approaches to impute body max index in a study of the relationship between osteoarthritis and cardiovascular disease. J Stat Res. 2020;54(2):131–45.
31. Desai RJ, et al. Evaluating the use of bootstrapping in cohort studies conducted with 1: 1 propensity score matching—A plasmode simulation study. Pharmacoepidemiol Drug Saf. 2019;28(6):879–86.
32. Bickel PJ, Sakov A. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. Statistica Sinica. 2008;18(3):967–85.
33. Gasparini A. Rsimsum: summarise results from Monte Carlo simulation studies. J Open Source Softw. 2018;3(26):739.
34. Gasparini A, White IR, Gasparini MA. Package 'rsimsum'; 2024.
35. Zhang Z. Too much covariates in a multivariable model may cause the problem of overfitting. J Thorac Dis. 2014;6(9).
36. Jacob D. Variable selection for causal inference via outcome-adaptive random forest. arXiv preprint arXiv:2109.04154; 2021.

Publisher's note