Arch Iran Med. April 2018;21(4):164-169

www.aimjournal.ir

ARCHIVES OF IRANIAN MEDICINE

Research Methods

Open Access

Causal Methods for Observational Research: A Primer

Amir Almasi-Hashiani, PhD Candidate[1]; Saharnaz Nedjat, MD, MPH, PhD[1,2]; Mohammad Ali Mansournia, MD, MPH, PhD[1*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[2]Knowledge Utilization Research Center, Tehran University of Medical Sciences, Tehran, Iran

Abstract

The goal of many observational studies is to estimate the causal effect of an exposure on an outcome after adjustment for confounders, but there are still some serious errors in adjusting confounders in clinical journals. Standard regression modeling (e.g., ordinary logistic regression) fails to estimate the average effect of exposure in total population in the presence of interaction between exposure and covariates, and also cannot adjust for time-varying confounding appropriately. Moreover, stepwise algorithms of the selection of confounders based on *P* values may miss important confounders and lead to bias in effect estimates. Causal methods overcome these limitations. We illustrate three causal methods including inverse-probability-of-treatment-weighting (IPTW) and parametric g-formula, with an emphasis on a clever combination of these 2 methods: targeted maximum likelihood estimation (TMLE) which enjoys a double-robust property against bias.

Keywords: Causal methods, Inverse-probability-of-treatment-weighting, Observational studies, Parametric g-formula, Targeted maximum likelihood estimation

Cite this article as: Almasi-Hashiani A, Nedjat S, Mansournia MA. Causal methods for observational research: a primer. Arch Iran Med. 2018;21(4):164–169.

Received: January 12, 2018, Accepted: March 4, 2018, ePublished: April 1, 2018

Introduction

The goal of many observational researches is to estimate the causal effect of a treatment (e.g. a specific surgical technique) or an exposure (e.g. alcohol consumption) on an outcome of interest (e.g. mortality). External variables (usually known as confounders) can distort the "causal" relationship. This phenomenon is called confounding.[1] The relationship between the exposure and outcome is confounded if the exposure effect is wholly or in part attributable to (or masked by) failure to control for confounders. Thus, confounders may exaggerate or weaken the causal relationship between exposure and outcome, or even may induce an apparent causal relationship when it does not really exist.[2] They should be associated with, but not affected by, both exposure and outcome, though this is not a sufficient condition.[3] Even randomized clinical trials are subject to random confounding i.e., randomization does not prevent confounding, but it makes confounding random.[4,5]

There are 2 broad approaches for controlling confounding at the analysis stage: standard regression modeling and causal methods. Standard regression methods are widely used in practice, but they suffer from some serious limitations. First, they cannot estimate the average causal effect of the exposure in presence of interaction between exposure and covariates, even if the target of the intervention on the exposure is the total population, and the interaction is not of interest.[6,7] The reason is that these methods assume no interaction between exposure and confounders to estimate the pooled effect, and in the presence of interaction, one has to report stratum-specific estimates. Moreover, the standard regression methods fail to appropriately adjust for time-varying confounding, and in fact, they cause over-adjustment and collider-stratification biases in the presence of time-varying confounding affected by prior treatment.[8–11] Causal methods could overcome these problems. The aim of this paper is to illustrate three causal methods including inverse-probability-of-treatment-weighting (IPTW) and parametric g-formula, with an emphasis on a clever combination of these 2 methods: targeted maximum likelihood estimation (TMLE).

Inverse-Probability-of-Treatment-Weighting

IPTW is based on propensity scores (PSs).[1,5,6,12] PS is the probability of exposure given the confounders. The weights used in the IPTW method are 1/PS in the exposed group and 1/(1 – PS) in the unexposed group.[2] The IPTW method has the following steps:

Step 1

Fit the exposure model i.e., the regression model of exposure on confounders using logistic regression or more advanced methods such as super learners (see below).

*Corresponding Author: Mohammad Ali Mansournia, MD, MPH, PhD; Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, P. O. Box: 14155-6446, Tehran, Iran. Email: mansournia_ma@yahoo.com

Step 2

Calculate the weight variable W from the exposure model fitted in step 1 as follows: W=1/PS for the exposed group and W=1/ (1 – PS) for the unexposed group.

Step 3

Calculate the weighted mean outcome (e.g. risk) in the exposed and unexposed groups using weights W calculated in step 2, and then calculate the effect measure of interest e.g., risk difference or risk ratio.

IPTW develops a pseudo-cohort, in which confounders do not predict the exposure and the causal effects are the same as in the real cohort. Thus the unadjusted (crude) analysis which is equivalent to the weighted analysis in real cohort, yields an unbiased estimate of the exposure effect assuming that there are no unmeasured confounders and PS does not equal 0 or 1.[13]

PSs are unknown and should be estimated, so that IPTW is based on exposure modeling i.e., the exposure is regressed on the confounders and then PS and the weights are estimated as mentioned in step 1 above. The method relies on the assumption that the exposure model is correctly specified. So, IPTW is especially recommended when the exposure mechanism is known. For example, in pharmacoepidemiologic studies, the predictors for exposure model are known using drug indications.[6] An important limitation of IPTW method is its sensitivity to huge weights of strong predictors of exposure among confounders which can introduce inefficiency and small-sample bias.[14] PSs can also be used in other procedures including PS matching (e.g., matching unexposed to the exposed based on PS) to estimate the causal effects. However most of these methods cannot be easily generalized to more realistic time-varying settings, and so we do not explain them here.[15–17]

Parametric g-Formula

Parametric g-formula is based on standard outcome regression modeling and standardization.[6] It is indeed a model-based generalization of classical standardization and is also known as model-based standardization.[18] Parametric g-formula has the following steps:

Step 1

Fit the outcome regression model i.e., the regression model of outcome on exposure and confounders. Logistic regression (for binary outcomes) or more advanced methods such as super learner can be used.

Step 2

Calculate the standardized mean outcome (e.g. risk) in the exposed group (A=1 where A denotes the exposure status taking 1 for the exposed and 0 for the unexposed groups) by predicting the individual mean outcome for exposure forced to be 1 for all individuals, and actual values of confounders, and then average them over the individuals from the model fitted in step 1. Similarly, calculate the standardized mean outcome (e.g. risk) in the unexposed group (A = 0) by predicting the individual mean outcome for exposure forced to be 0 for all individuals, and actual values of confounders, and then average them over the individuals from the model fitted in step 1. Then calculate the effect measure of interest e.g., risk difference or risk ratio.

Parametric g-formula relies on the correct specification of the outcome model. It is preferred to IPTW if we are confident about the correctness of the outcome model than exposure model, or if there are huge inverse-probability-of-treatment weights. Another consideration for choosing between IPTW and parametric g-formula is the relative frequency of exposure and outcome which may affect sparse-data bias. If exposure is more common than outcome, IPTW is preferred, and if outcome is more common than exposure, parametric g-formula is preferred.[8,19–21]

Targeted Maximum Likelihood Estimation

IPTW and parametric g-formula can be combined in various ways to produce a double robust causal method in the sense that it has 2 possibilities to adjust for confounders i.e., the method provides unbiased estimate of causal effect if either exposure model or outcome model is correct. One possibility is TMLE. In brief, TMLE first estimates the outcome regression model i.e., conditional expectation of the outcome given exposure and confounders. Then, the initial estimate is updated using a new covariate that reduces the bias of the initial estimate for the target causal parameter. The covariate gets information from the exposure model i.e., the probability of the exposure given confounders and intended to reduce bias.[22-24] The TMLE has the following steps:

Step 1

Fit the outcome regression model i.e., the regression model of outcome on exposure and confounders. Logistic regression (for binary outcomes) or more advanced methods such as super learner can be used.

Step 2

Fit the exposure model i.e., the regression model of exposure on confounders using logistic regression or more advanced methods such as super learners.

Step 3

Calculate the variable H from the exposure model fitted

in step 2 as follows:
H = A/PS – ((1-A) /(1 – PS)).

Step 4

Refit the outcome regression model from step 1 by adding the variable H, so that the coefficients of the model do not change. This can be done by declaring the right hand side of the outcome model fitted in step 1 as "offset" (i.e., a variable whose coefficient is forced to be 1) in the new model and suppressing the constant. Note that variable H equals 1/PS=W for the exposed and -1/(1-PS) = -W for the unexposed groups.

Step 5

Calculate the standardized mean outcome (e.g. risk) in the exposed group (A=1) by predicting the individual mean outcome, for exposure forced to be 1 for all individuals, and actual values of confounders, and then average them over the individuals from the model fitted in step 4. Similarly, calculate the standardized mean outcome (e.g. risk) in the unexposed group (A=0) by predicting the individual mean outcome, for exposure forced to be 0 for all individuals, and actual values of confounders, and then average them over the individuals from the model fitted in step 4.[22–25] Then calculate the effect measure of interest e.g., risk difference or risk ratio.

The steps mentioned above suggest that TMLE is an extension of parametric g-formula in which information from IPTW is used, because the variable H is a function of inverse-probability-of-treatment weights for the exposed (1/PS) and the (1/(1-PS)) groups. Statistical theory tells us that the variable H changes the initial estimate in the direction of the target causal parameter, optimizing the bias-variance trade-off for the parameter of interest. Under the assumption of no unmeasured confounding, TMLE provides unbiased estimates of causal effects if either outcome regression model fitted in step 1 or exposure regression model fitted in step 2 is correct. A recent simulation study[25] suggested that TMLE performs well even in the presence of significant model misspecifications such as an omitted confounder in either exposure or outcome model. Moreover, TMLE is efficient if both of these models are correctly specified. It is generally more efficient than g-formula for the parameter of interest, as g-formula optimizes the bias-variance trade-off for the outcome regression model, but not the causal parameter of interest.

There are other double-robust causal methods like TMLE which combine outcome and exposure models e.g., weighted parametric g-formula (where the weights are inverse-probability-of-treatment weights), augmented IPTW (which augments IPTW with a term in the outcome model), etc. However, TMLE has shown to be superior to these double-robust methods especially in terms of robustness to outliers and data sparsity.[22,26]

Confidence intervals for all causal methods can be derived using non-parametric bootstrapping in which K random samples of size N (the size of the study sample) are taken from the data set by sampling with replacement. All steps of the causal method are repeated in each sample to obtain k causal effect estimates. Confidence intervals can be derived using 2.5 percentile and 97.5 percentile values of these K estimates. Another common method for deriving confidence intervals is using influence function IF which assesses changes in causal effect estimator by addition of each person's profile (i.e., exposure, outcome, and confounders). The standard error for the causal effect estimator equals the square root of the estimate of the sample variance of IF (which is calculated for each person) divided by the sample size. Confidence intervals and *P* values can be calculated using this estimate of standard error. Finally, robust "sandwich" variance estimators based on the observed variability in data, can be used to provide valid confidence intervals for IPTW and parametric g-formula methods.[2]

We have illustrated three causal methods for time-fixed cohort studies, but they also can be easily generalized to handle time-varying confounding in longitudinal studies.[27] Moreover, all of these methods including IPTW, parametric g-formula and TMLE can be used for case-control studies, but adjustments are needed to account for different sampling fractions of cases and controls. Specifically, TMLE can be extended for case-control studies using case-control weighted TMLE in which all steps 1-2 and 4-5 mentioned above are weighted with weights equal to the outcome prevalence for cases and ((1- outcome prevalence)/control-to-case sampling ratio) for controls.[23,28] Other generalizations include polychotomous and continuous exposures[9] instead of binary exposures, as well as impact measures such as population attributable fraction instead of effect measures.[29,30]

Model Specification

As explained before, all causal methods rely on the correct specification of the models. In particular, IPTW and parametric g-formula require correct specification of the exposure and outcome models, respectively. For TMLE, either exposure or outcome model needs to be correct. Correct specification of regression model requires including minimal sufficient set of confounders, choosing appropriate scale for continuous variables, and inclusion of interaction terms (product terms) between predictors (exposure and confounders) if needed. Since data alone say nothing about confounding, minimal

sufficient set of confounders should be determined based on subject matter knowledge. They can be conveniently derived using causal directed acyclic graphs (causal diagrams).[15,31] Change-in-estimate criterion can be used to exclude unimportant confounders in stepwise manner i.e., if the difference in adjusted and unadjusted estimate divided by adjusted estimate is less than 10%.[32] However, we note that stepwise algorithms of the selection of confounders based on P-values which is often used in practice, may miss important confounders and lead to bias in effect estimates, so they should not be used for the effect estimation.

There are several modeling approaches to account for complex scales and relationships between variables including neural network, Bayesian methods, classification and regression trees, random forests, generalized boosted regression, and generalized additive models.[22–24,33] Alternatively, fractional polynomial[34] and spline regression[35] can be used to choose appropriate scales of continuous variables in standard regression models such as logistic regression.

A very powerful method for exposure and outcome modeling in causal inference is super learning. It is a machine learning approach based on a linear combination of several regression modeling approaches such as logistic regression, neural network, Bayesian approach, classification and regression trees, random forests, generalized boosted regression, generalized additive models, etc. The optimal weight for contribution of each model is derived so that the mean squared error (i.e., the mean of squared difference between observed value of outcome and predicted value of outcome by the model) using k-fold cross-validation procedure is minimized.[36] The k-fold cross-validation procedure involves partitioning the sample into k subsamples and fitting the data on (k-1) subsamples ("training set") and evaluating the fit in the remaining subsample ("validation set"). The important point is that performance of super learning in terms of cross-validated mean squared error is better than each included model.

Software

The causal methods discussed in this paper can be performed using general softwares along with the steps motioned above. Alternatively, one can use special commands in Stata. For example, IPTW can be done using teffects ipw command and parametric g-formula can be fitted using margins command in Stata. TMLE along with super learner can be implemented using eltmle stata command as well as tmle R package. All analyses in the case study (see below) were performed using Stata software, version 13 (Stata Corp, College Station, TX, USA) and R software (version 3.4.3).

Case Study

To illustrate the use of TMLE for real data, we estimated the effect of preeclampsia on the preterm delivery from a cohort of 3000 pregnant women who were referred to 103 hospitals for delivery in Tehran province, 2014. Preeclampsia was defined as the presence of high blood pressure (a systolic blood pressure ≥140 mm Hg or a diastolic blood pressure ≥90 mm Hg after 20 weeks of gestational age) and proteinuria. Preterm delivery was defined based on the World Health Organization (WHO) as the live birth of a newborn before 37 weeks of pregnancy are completed.

The potential confounders included parents' age, mother's education, mother's occupation, household economic status, pre-pregnancy body mass index (BMI), unwanted pregnancy, number of previous pregnancies, number of previous deliveries, cesarean section, low birth weight, maternal weight gain during pregnancy, history of previous abortion(s), and the number of previous abortion(s).

We estimated the effect of preeclampsia on preterm delivery using TMLE as the causal method and super learning for model specification. *P* values and 95% CIs were derived using influence function (IF) methodology. The default super learner algorithms implemented in tmle R package were used: (*i*) SL.step: stepwise forward and backward model selection using Akaike information criterion (AIC), restricted to quadratic polynomials, (*ii*) SL.glm: ordinary logistic regression without interaction terms or polynomials, and (*iii*) SL.glm.interaction: a variant of logistic regression model that includes quadratic polynomials and two-by-two interactions of the main terms included in the model. Using TMLE and super learning, the causal risk ratio and risk difference estimates were 1.64 (95% CI: 1.13– 2.39, $P = 0.009$), and 0.055 (95% confidence interval: 0.005, 0.105, $P = 0.032$), respectively. The software also reported an estimate of odds ratio (95% CI) as 1.75 (1.13–2.69). However, we warn that odds ratio is not a desirable causal effect measure as it overestimates the risk ratio and also suffers from a mathematical peculiarity known as non-collapsibility when the outcome is not uncommon (as is the case in our example).[37-40]

The contribution of 3 algorithms of SL.glm, SL.step and SL.glm.interaction in super learner for exposure and outcome modeling in TMLE are shown in Table 1. It highlights that including interaction and quadratic terms improves the accuracy of both models, but stepwise selection has nothing to add.

Discussion

The goal of many observational studies is to estimate the causal effect of an exposure on an outcome after

Table 1. The Contribution (Weights) of 3 Algorithms in Super Learner for Exposure and Outcome Modeling

Algorithms	Exposure Model	Outcome Model
SL.glm	0.86	0.74
SL.step	0	0
SL.glm.interaction	0.14	0.26

SL.step, stepwise forward and backward model selection using Akaike information criterion (AIC), restricted to quadratic polynomials. SL.glm, ordinary logistic regression without interaction terms or polynomials. SL.glm.interaction, a variant of logistic regression model that includes quadratic polynomials and two-by-two interactions of the main terms included in the model.

adjustment for confounders, but there are still errors in adjusting confounders in clinical journals: (*i*) selection of confounders is often based on associations in data at hand and selection algorithms like stepwise regression which may miss important confounders and lead to bias in effect estimates. (*ii*) Clinical researchers often use conditional methods of adjustment such as regression modeling, though these methods fail to estimate the average effect of exposure in total population in the presence of interaction(s) between exposure and covariates, and also cannot appropriately adjust for time-varying confounding. We encourage clinical researchers to use causal methods for causal effects estimation. Ideally, a double-robust causal method such as TMLE should be used along with super learner to avoid model misspecification.

Authors' Contribution

AA and MAM jointly wrote the text and SN suggested critical revisions. All authors read and approved the final manuscript.

Conflict of Interest Disclosures

The authors have no conflicts of interest.

Ethical Statement

Not applicable.

References

1. Suzuki E, Tsuda T, Mitsuhashi T, Mansournia MA, Yamamoto E. Errors in causal inference: an organizational schema for systematic error and random error. Ann Epidemiol. 2016;26(11):788-93. e1. doi: 10.1016/j.annepidem.2016.09.008.
2. Mansournia MA, Altman DG. Inverse probability weighting. Bmj. 2016;352. doi: 10.1136/bmj.i189.
3. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2008:128-47.
4. Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. Eur J Epidemiol. 2015;30(10):1101-10. doi: 10.1007/s10654-015-9995-7.
5. Mansournia MA, Altman DG. Invited commentary: methodological issues in the design and analysis of randomised trials. Br J Sports Med. 2018;52(9):553-5. doi: 10.1136/bjsports-2017-098245.
6. Gharibzadeh S, Mohammad K, Rahimiforoushani A, Amouzegar A, Mansournia MA. Standardization as a Tool for Causal Inference in Medical Research. Arch Iran Med. 2016;19(9):666-70. doi: 0161909/aim.0011.
7. Mohammad K, Hashemi-Nazari SS, Mansournia N, Mansournia M. Marginal versus conditional causal effects. J Biostat Epidemiol. 2015;1(3-4):121-8.
8. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. Bmj. 2017;359:j4587. doi: 10.1136/bmj.j4587.
9. Mansournia MA, Danaei G, Forouzanfar MH, Mahmoodi M, Jamali M, Mansournia N, et al. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis : analysis with marginal structural models. Epidemiology. 2012;23(4):631-40. doi: 10.1097/EDE.0b013e31824cc1c3.
10. Shakiba M, Mansournia MA, Salari A, Soori H, Mansournia N, Kaufman JS. Accounting for Time-varying Confounding in the Relation between Obesity and Coronary Heart Disease: Analysis with G-estimation, the Atherosclerosis Risk in Communities (ARIC) study. Am J Epidemiol. 2017. doi: 10.1093/aje/kwx360.
11. Salimi Y, Fotouhi A, Mohammad K, Mansournia N, Mansournia MA. Causal Effects of Intensive Lifestyle and Metformin Interventions on Cardiovascular Disease Risk Factors in Pre-Diabetic People: An Application of G-Estimation. Arch Iran Med. 2017;20(1):55-9. doi: 0172001/aim.0012.
12. Almasi-Hashiani A, Mansournia MA, Rezaeifard AR, Mohammad K. Causal effect of donor source on survival of renal transplantation using marginal structural models. Iran J Public Health. 2018; In Press.
13. Hernán MA, Robins JM. Causal Inference: Boca Raton: Chapman & Hall/CRC, forthcoming; 2018.
14. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21(1):31-54. doi: 10.1177/0962280210386207.
15. Mansournia MA, Hernan MA, Greenland S. Matched designs and causal diagrams. Int J Epidemiol. 2013;42(3):860-9. doi: 10.1093/ije/dyt083.
16. Mansournia MA, Jewell NP, Greenland S. Case-control matching: effects, misconceptions, and recommendations. Eur J Epidemiol. 2018;33(1):5-14. doi: 10.1007/s10654-017-0325-0.
17. Shinozaki T, Mansournia MA, Matsuyama Y. On hazard ratio estimators by proportional hazards models in matched-pair cohort studies. Emerg Themes Epidemiol. 2017;14:6. doi: 10.1186/s12982-017-0060-8.
18. Gharibzadeh S, Mansournia MA, Rahimiforoushani A, Alizadeh A, Amouzegar A, Mehrabani-Zeinabad K, et al. Comparing different propensity score estimation methods for estimating the marginal causal effect through standardization to propensity scores. Commun Stat Simul Comput. 2017:1-13. doi: 10.1080/03610918.2017.1300267.
19. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. Bmj. 2016;352:i1981. doi: 10.1136/bmj.i1981.
20. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in Logistic Regression: Causes, Consequences, and Control. Am J Epidemiol. 2018;187(4):864-70. doi: 10.1093/aje/kwx299.
21. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat Med. 2015;34(23):3133-43. doi: 10.1002/sim.6537.
22. van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer Science & Business Media; 2011.
23. Rose S, van der Laan M. A double robust approach to causal effects in case-control studies. Am J Epidemiol. 2014;179(6):663-9. doi: 10.1093/aje/kwt318.
24. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017;185(1):65-73. doi: 10.1093/aje/kww165.
25. Rose S. Causal Inference for Case-Control Studies. Berkeley: University of California; 2011.

26. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. Int J Biostat. 2010;6(1):Article 17. doi: 10.2202/1557-4679.1181.
27. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. Bmj. 2017;359:j4587. doi: 10.1136/bmj.j4587.
28. Abdollahpour I, Nedjat S, Mansournia MA, Sahraian MA, Kaufman JS. Estimating the Marginal Causal Effect of Fish Consumption during Adolescence on Multiple Sclerosis: A Population – Based Incident Case – Control Study. Neuroepidemiology. 2018; In Press.
29. Fleischer NL, Fernald LC, Hubbard AE. Estimating the potential impacts of intervention from observational data: methods for estimating causal attributable risk in a cross-sectional analysis of depressive symptoms in Latin America. J Epidemiol Community Health. 2010;64(1):16-21. doi: 10.1136/jech.2008.085985.
30. Mansournia MA, Altman DG. Population attributable fraction. Bmj. 2018;360:k757. doi: 10.1136/bmj.k757.
31. Mansournia MA, Higgins JP, Sterne JA, Hernan MA. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. Epidemiology. 2017;28(1):54-9. doi: 10.1097/ede.0000000000000564.
32. Greenland S, Pearce N. Statistical foundations for model-based adjustments. Annu Rev Public Health. 2015;36:89-108. doi: 10.1146/annurev-publhealth-031914-122559.
33. Dreyfus G. Neural networks: methodology and applications. Springer Science & Business Media; 2005.
34. Royston P, Sauerbrei W. Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley & Sons; 2008.
35. Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2015.
36. Rose S. Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol. 2013;177(5):443-52. doi: 10.1093/aje/kws241.
37. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology. 2015;26(4):466-72. doi: 10.1097/ede.0000000000000291.
38. Janani L, Mansournia MA, Nourijeylani K, Mahmoodi M, Mohammad K. Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies. Arch Iran Med. 2015;18(10):713-9. doi: 0151810/aim.0012.
39. Diaz-Quijano FA, Janani L, Mansournia MA. Cluster vs. Robust Estimation of Risk Ratio using Expanded Logistic Regression. Arch Iran Med. 2016;19(8):608-9. doi: 0161908/aim.0015.
40. Janani L, Mansournia MA, Mohammad K, Mahmoodi M, Mehrabani K, Nourijelyani K. Comparison between Bayesian approach and frequentist methods for estimating relative risk in randomized controlled trials: a simulation study. J Stat Comput Simul. 2017;87(4):640-51. doi: 10.1080/00949655.2016.1222610.