Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2017; 3(3-4): 76-82.

Original Article

Generation of data with specific marginal risk difference

Kazem Mohammad[1], Mohammad Ali Mansournia[2], Safoora Gharibzadeh[3*]

[1] Professor, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[2] Assistant Professor, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[3] Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Received 25.03.2017
Revised 19.06.2017
Accepted 23.07.2017
Published 01.10.2017

Key words:
Data systems;
Risk ratio;
Causality;
Computer simulation;
Monte Carlo method

ABSTRACT

Background & Aim: Simulation studies are important statistical tools to investigate the performance of statistical models in specific situations. For a binary outcome and exposure, one of the most important statistical measures will be the risk difference (RD). To assess the quality of estimators in estimating the effect of the exposure, a data set with a specific effect measure is require.

Methods & Materials: Monte Carlo simulation can be helpful in situations when there is a proper data generating process. In this paper, another technique will be presented to generate data with specific marginal risk difference (MRD).

Results: Convergence of simulation methods in the same scenario reached in a few iterations using the proposed method.

Conclusion: The proposed method is recommended over the current method due to less time consumption; this issue is important in studies with different scenarios.

Introduction

Estimating the causal effects of exposure using observational data is a common problem in medical research (1, 2). In ideal randomized experiments, association measures can be interpreted causally as randomization ensures that the exposed and the unexposed are exchangeable. In observational studies, however, association does not ensure causation; association measures cannot be interpreted causally since the exposed and the unexposed are not generally exchangeable (3).

When randomization is not feasible and observational data are required to be used to estimate marginal treatment effect, one of the most important measures of association is risk difference (RD) and number needed to treat (1).

Monte Carlo simulation is an important tool in modern statistical methods. Simulation methods allow researchers to investigate the efficiency of estimators in settings in which mathematical derivations are difficult. However, the use of Monte Carlo simulation relies on the existence of well-suited data-generating processes.

With the binary outcome and exposure, it is a difficult task to generate a data with a specific marginal risk difference (MRD) from a conditional data-generating process (4).

Several methods have been proposed for different study designs and complicated statistical models. As an example, Austin and Stafford introduced a method for generating data

* Corresponding Author: Safoora Gharibzadeh, Postal Address: Email: safoora.gharibzadeh@gmail.com

Please cite this article in press as: Mohammad K, Mansournia MA, Gharibzadeh S. Generation of data with specific marginal risk difference. *J Biostat Epidemiol. 2017; 3(3-4): 76-82*

with specific marginal odds ratio (5). Austin proposed a method for data generation with specific MRD or number needed to treat (1). Leemis et al. explained a method to generate data for accelerated life and proportional hazards models with time-dependent covariates (4). Bender et al. proposed a method to simulate survival times for a Cox proportional hazards model (6). Lunn and Davies proposed a method for generating correlated binary variables (7), and recently, Austin proposed a method for generating survival times in Cox proportional hazards models with time-varying covariates (8).

Based on literature review performed by the researchers in the present study, only the method proposed by Austin generated the data with specific MRD (1). Generating a scenario in this method takes almost 1 hour. In case of diversity of scenarios, using this method is time-consuming and needs iterative Monte Carlo simulations.

The objective of the current study was to present a method with less time consumption and lacking the need for iterative procedures. The remainder of this article is organized as follows. Section 2, reviews the method proposed by Austin (1). Section 3 presents notation used in the data generating process. Section 4 demonstrates the new proposed method using an example. Finally, Section 5 summarizes the results and concludes with a discussion of the findings.

Methods

Review of the current method: Austin (1) proposed a method for generating data with specific MRD. This method will be described briefly in the following.

Based on the known relationship between exposure and outcome, it was found that the conditional effect of treatment (β) induced a specific MRD.

Using counterfactual framework, two potential outcomes were assumed for each subject, and then the probability of these counterfactual outcomes (all subjects were considered treated and untreated simultaneously) were defined; thus:

$$\log it\big(\pi(y=1)\big) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p + \beta.\mathrm{Tr}$$

$$p_{i1} = \frac{1}{1+\exp(-(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}+\cdots+\alpha_p x_{pi}+\beta}$$

$$p_{i0} = \frac{1}{1+\exp(-(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}+\cdots+\alpha_p x_{pi}}$$

Austin (1) has defined the marginal probability of success as:

$$\bar{p}_1 = \int_{x_p}\dots\int_{x_2}\int_{x_1}\frac{1}{1+\exp\left(-(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}\dots+\alpha_p x_{pi}+B)\right)} f(x_1)f(x_2)\dots f(x_p)d_{x1}d_{x2}\dots d_{xp}$$

$$\bar{p}_0 = \int_{x_p}\dots\int_{x_2}\int_{x_1}\frac{1}{1+\exp\left(-(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}\dots+\alpha_p x_{pi})\right)} f(x_1)f(x_2)\dots f(x_p)d_{x1}d_{x2}\dots d_{xp}$$

The method was based on iterative computation of the MRD for specific values of β using Monte Carlo integration. The iterative process allows selecting the value of β which induces a MRD. Suppose that β (k) denotes the value of β at the k[th] iteration. Using p, explanatory variables $X_1, X_2, \dots, X_p$ were generated randomly for each of n subjects from a specified distribution. At k[th] step of Monte Carlo simulation, the probability of success in case of treatment was computed as follows

$$p_{i,1} = \frac{1}{1+\exp\left(-\left(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}\dots+\alpha_p x_{pi}+\beta\right)\right)}$$

And if untreated:

$$p_{i,0} = \frac{1}{1+\exp\left(-\left(\alpha_0+\alpha_1 x_1+\alpha_2 x_2\dots+\alpha_p x_p\right)\right)}$$

Then, the mean of each of these two probabilities across the simulated data should be computed:

$$\bar{p}_0^{mc(k)} = \frac{1}{n}\sum_{i=1}^{n} p_{i,0}$$

$$\bar{p}_1^{mc(k)} = \frac{1}{n}\sum_{i=1}^{n} p_{i,1}$$

In the next step, the empirical MRD was calculated as:

$$\gamma_{(n)}^{(k)} = \bar{p}_0^{mc(k)} - \bar{p}_1^{mc(k)}$$

The empirical MRD $\gamma_{(n)}^{(k)}$ represents the MRD

in the n^{th} randomly generated data set at the k^{th} step of the iterative process. This process is then repeated across 1000 simulated data sets and the mean MRD is determined as $\gamma^{(k)} = \frac{1}{1000}\sum_{n=1}^{1000}\gamma_{(k)}^{(n)}$. The quantity $\gamma^{(k)}$ represents the empirical MRD after the k^{th} step of the iterative process. To calculate β, Austin (1) used bisection method to determine a solution to an equation. It was reported in this study that computations for each sensible scenario required fewer than 52 minutes in R Version 2.8.0 (R Core Development Team, 2005).

In the present study, RD was considered as a measure of association. RD is defines as a difference between marginal probability of outcome in treated and untreated subjects.

Changes in the marginal probability: Suppose t p confounders $X_1, X_2, Z_1, \dots, Z_q$. Y = 1 denotes success. If π denoted the probability of success, the relationship between the confounders, treatment status, and the logit of the probability of outcome could be described as below:

$$\text{logit}\big(\pi(y=1)\big) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_p Z_p + \beta Tr \quad (1)$$

X_i and Z_i were categorical and continuous variables, respectively, and Tr was treatment status. Probability of outcome in treated and untreated subjects could be defined as below:

$$p_{i,1} = \frac{1}{1+e^{-(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_p x_p+\cdots+\beta)}} \quad (2)$$

$$p_{i,0} = \frac{1}{1+e^{-(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\alpha_p x_p)}} \quad (3)$$

Using the joint probability of the categorical variables in the present study, the population could be reconstructed to new subgroups. For example, in case of having one binary variable (X_1) (with the coefficient of α_1), two new sets could be defined with logit of the baseline probability equal to α_0 and $(\alpha_0 + \alpha_1)$. α_0 and $(\alpha_0 + \alpha_1)$ were called as "constant part" of the model. In other words, different categories could affect the baseline probability (or the intercept term of the model).

$$\text{logit}\big(\pi(y=1)\big) = \alpha_0 + \alpha_1 x_1 + \beta \text{Tr}$$
$$\begin{cases} \text{logit}\,\big(\pi(y=1)\big) = (\alpha_0 + \alpha_0) + \beta \text{Tr}, x = 1 \\ \text{logit}\,\big(\pi(y=1)\big) = \alpha_0 + \beta \text{Tr}, x = 0 \end{cases} \quad (4)$$

For k standard normal variables, the sum of these k independent normal variables was normal with variance equal to λ:

$$Z_i \sim N(0,1): Z = \sum_{i=1}^{k} \alpha_i Z_i \sim N(0, \lambda = \sum \alpha_1^2) \quad (5)$$

All of the continuous variables could be summarized into one component; this part of the model (Z) was called as "random part".

According to the above paragraphs, for each combination of categorical covariates, the equation (1) could be rewritten as:

$$\text{logit}\big(\pi(y=1)\big) = \text{Constant part} + \text{Random part} + \beta.\text{Tr} \quad (6)$$

Adding categorical variable to the model changed the baseline prevalence; in other words, a shift emerged in the intercept, so the marginal probability could be easily computed in treated and untreated subjects.

When the model contained continuous variables, compute marginal probability had to be computed using Monte Carlo integration. Adding a continuous variable into the model could be considered as adding a random part into the model.

In this part, it was shown that what changes will happen in the marginal probability by adding continuous variables into the model.

The effect of different levels of the random part was evaluated in altering the marginal probability. The variation in the random part (continuous variable) was set 0 to 3 with 0.01 increments, and simultaneously, the constant part was changed from 0.0 to 1.0 with 0.0001 increments. Table 1, which is shown partially below, can be found in the appendix. Each row in this table corresponds to one of the baseline probabilities, and columns show the related marginal probability corresponding to particular random part (continuous variables).

In figure 1, the graphs were plotted for the random parts equal to 1, 2…, 7. Each of the curves corresponds to the probabilities of one of the random parts, and the red curve, which lies on the bisector of the first and third quadrants, represents the original values.

In the presence of the random part with variance of λ, different marginal probabilities existed for a fixed baseline probability (p).

Table 1. Different values of marginal probability corresponding to different random parts

Fixed part (baseline P)	Random part (λ)								
	0.1	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**5.0**	**7.0**
0.05	0.0521	0.0611	0.0727	0.0842	0.0954	0.1057	0.1158	0.1500	0.1770
0.10	0.1035	0.1176	0.1340	0.1485	0.1617	0.1736	0.1847	0.2190	0.2441
0.15	0.1543	0.1708	0.1884	0.2033	0.2168	0.2283	0.2385	0.2700	0.2922
0.20	0.2046	0.2214	0.2385	0.2529	0.2651	0.2756	0.2845	0.3120	0.3310
0.25	0.2545	0.2702	0.2859	0.2983	0.3089	0.3185	0.3259	0.3485	0.3640
0.30	0.3039	0.3176	0.3312	0.3414	0.3497	0.3579	0.3638	0.3819	0.3946
0.35	0.3532	0.3643	0.3748	0.3827	0.3894	0.3945	0.3995	0.4132	0.4225
0.40	0.4022	0.4101	0.4165	0.4226	0.4266	0.4303	0.4332	0.4429	0.4490
0.45	0.4512	0.4549	0.4584	0.4612	0.4639	0.4659	0.4672	0.4169	0.4748
0.50	0.4999	0.5000	0.4994	0.5000	0.5004	0.4996	0.4996	0.4999	0.4999
0.55	0.5488	0.5448	0.5415	0.5387	0.5368	0.5348	0.5326	0.5283	0.5253
0.60	0.5976	0.5899	0.5831	0.5776	0.5733	0.5696	0.5666	0.5569	0.5509
0.65	0.6467	0.6356	0.6253	0.6173	0.6110	0.6055	0.6005	0.5868	0.5775
0.70	0.6959	0.6822	0.6686	0.6581	0.6497	0.6428	0.6365	0.6179	0.6056
0.75	0.7454	0.7296	0.7139	0.7011	0.6911	0.6823	0.6742	0.6512	0.6356
0.80	0.7952	0.7784	0.7612	0.7471	0.7343	0.7249	0.7156	0.6879	0.6690
0.85	0.8450	0.8292	0.8116	0.7962	0.7830	0.7717	0.7609	0.7300	0.7078
0.90	0.8964	0.8824	0.8659	0.8516	0.8382	0.8262	0.8155	0.7808	0.7556
0.95	0.9478	0.9389	0.9273	0.9157	0.9045	0.8940	0.8841	0.8498	0.8228

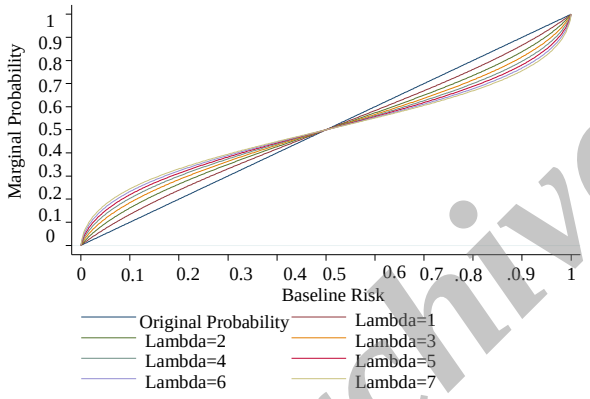

Figure 1. Baseline and marginal probability

For example, with random part of *3* and baseline probability of 0.2, the marginal probability of outcome was equal to 0.2847. It meant that the probability of outcome changed from 0.2000 to 0.2847, and if the random part equaled 10, the marginal probability reached 0.4455.

Increasing the random part, the probability approached 0.5. Changes in probabilities were in the form of centralizing, meaning converting the baseline probability into 0.5. Based on figure 1, the rate of convergence increases with an increase in the random part.

Changes in the population probabilities: In the presence of the random part, which value of baseline probability can be reached in case of having a definite RD?

In the potential outcome framework, each subject has two counterfactual outcomes: one for Tr = 1 and one for Tr = 0 ($Y^{(1)}, Y^{(0)}$); thus, two probabilities can be obtained as:

$$p_1 = \text{expit}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p z_p + \beta) \quad (7)$$

$$p_0 = \text{expit}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p z_p) \quad (8)$$

And MRD can be defined as:

$$R.D = \overline{P}_1 - \overline{P}_0 (6)$$

In case of two binary variables and k continuous variables, as noted earlier, the population can be reconstructed into four subpopulations with different baseline probabilities. Now, the effect of the random part on the whole population will be shown using table 1.

The values in the second column of table 2 are baseline probabilities in each of the subpopulations. After the inclusion of continuous variables in the model, calculating the probability of each subject, and averaging, the marginal probability would be included in untreated subjects, shown in column 3.

Table 2. Baseline and marginal probabilities in treated and untreated subpopulations

Subpopulation	Baseline P untreated	Marginal P untreated	Marginal P treated	Baseline P treated
$x_1 = x_2 = 0$	p1	P_1'	P_1'+RD	P_1''
$x_1 = o, x_2 = 1$	p2	P_2'	P_2' + RD	P_2''
$x_1 = 1, x_2 = 0$	p3	P_3'	P_3' + RD	P_3''
$x_1 = 1, x_2 = 1$	p4	P_4'	P_4'+ RD	P_4''

By adding the desired RD to marginal probability of untreated subjects, the marginal probability among treated subjects will be as column 4. According to table 1, one can check which values of baseline probabilities (column 5) correspond to these marginal probabilities.

For example, when random part equals zero (Z = 0) in the first sub-population, after re-writing equations (4) and (5), we have:

$$\text{untreated: } \text{logit}(p_0) = \alpha_0 \quad (10)$$

$$\text{treated: } \text{logit}(p_0'') = \alpha_0 + \beta \quad (11)$$

$$\beta = \text{logit}(p_0'') - \text{logit}(p_0) \quad (12)$$

According to table 3, β can be computed for each sub-population.

Table 3. Logit (p) of baseline probability in different subpopulations

Subpopulation	Untreated	Treated
$x_1 = x_2 = 0$	α_0	$\alpha_1' = \alpha_0 + \beta_1$
$x_1 = o, x_2 = 1$	$\alpha_0 + \alpha_2$	$\alpha_2' = \alpha_0 + \alpha_2 + \beta_2$
$x_1 = 1, x_2 = 0$	$\alpha_0 + \alpha_1$	$\alpha_3' = \alpha_0 + \alpha_1 + \beta_3$
$x_1 = 1, x_2 = 1$	$\alpha_0 + \alpha_{1+} \alpha_2$	$\alpha_4' = \alpha_0 + \alpha_1 + \alpha_2 + \beta_4$

To generate a data set with desired marginal RD, one can:

1. according to collapsibility of RDs (9), generate data in a way that each part of the data is generated with its own β.
2. compute total β through one of these methods:

a. Computing total β as a weighted combination of these βs.

After computing β for each part of the population, one can assume that the total β is in the range of these βs. Then, the total β can be found using bisection method.

Results

The above algorithm was explained with an example:

Two binary variables with marginal probabilities were present equal to 0.2 and 0.5, and the joint probability as bellow:

	0	1	
0	0.15	0.05	0.2
1	0.35	0.45	
	0.5		

In case of generating a data set with RD = +0.05 with baseline probability of outcome equal to 0.15, two binary variables with parameters 0.2 and 0.5, OR = 2, and 4 standard normal variables with OR equal to 2.0, 1.5, 2.0, and 3.0, β will be computed as below (Table 4):

$$\text{logit}\big(\pi(y = 1)\big) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta.\text{Tr}$$

$$\lambda = \sum \beta_i^2 \sigma^2 = ((\log 2)^2 + (\log 1.5)^2 + (\log 2)^2 + (\log 3)^2)$$

$$\lambda = 2.232$$

$$\alpha_0 = \log(0.15/0.85) = -1.735$$

A comparison of RD with corresponding β is summarized in table 5. In this example, the second method (bisection method for calculating β) yielded the best result compared to the other methods.

Table 4. Baseline and marginal probabilities in treated and untreated subpopulations in the example

Proportion in population	Baseline probability of untreated (p0)	Marginal probability (λ = 2.33)	Baseline probability of treated (P_0')	βi
0.15	0.1500	0.2245	0.2025	0.3638
0.05	0.2610	0.3236	0.3238	0.3044
0.35	0.2610	0.3236	0.3238	0.3044
0.45	0.4138	0.4389	0.4845	0.1285

Table 5. Estimates of β and calculated risk difference (RD)

Method	Estimated β	RD
1	-	0.05440
2.a	0.2342	0.03540
2.b*	0.3270	0.04995
Austin	0.3270	0.04999

*After 5 iterations

Discussion

The objective in this article was to describe a data-generating process for binary outcomes with a specific MRD. This method was based on the changes in baseline probability of the outcome due to the entrance of continuous random variables and binary treatment.

In brief, data sets were generated with different final normal distribution (sum of various normal distributions) and the marginal probability for each condition was computed with regard to different baseline probabilities varying from 0 to 1 with 0.0001 increments (the supplementary file). Therefore, according to the joint distribution of categorical variables and using probability table, one can find the coefficient of treatment inducing the desired MRD in each subpopulation and then compute total β.

Although the logic of both methods is the same and based on the changes in the marginal probability in treated and untreated subjects, the method presented in this study outperformed the method proposed by Austin (1) due to shorter time to reach the coefficient of exposure/treatment and lack of requiring iterative methods. This method can be very helpful in case of presence of many different scenarios.

In the method presented in this study, adding continuous variables imposes not a special problem in simulation process. These variables affect the marginal probability through their variances, and the effect of these variables can be assessed using the sum of them as a single new variable.

Another interesting finding is that, in case of changing the fixed part from p to (1-p), the same change can be observed, but in the opposite direction. For instance, when the random part equals 3, considering a baseline probability of 0.88 (1-p) instead of 0.12, then a marginal probability of 0.7927 will be achieved. With some rounding error, the difference between this value and 0.5 is equal to the difference between 0.2073 and 0.5.

When baseline risk (p_0) is changed to $1-p_0$, the sum of two marginal probabilities will be equal to 1. By changing the random part, the difference of both from 0.5 will be the same. According to figure 1, when the random part increases, both values approach 0.5. In small fixed part (close to zero), larger random parts are needed to achieve 0.5. If p_0 was around zero, all the values would approach 0.5 and the dominant part would be the random part.

In the hypothetical scenario in this study, β equals 0.3270 and 32703 using the Austin method (1) and the bisection method of this study, respectively. In order to calculate the coefficient of treatment, Austin's method used iterative Monte Carlo method, which is time-consuming and needs generating data for each scenario.

With this method, the problem of using Monte Carlo integration was solved in order to calculate marginal probabilities, and in case of large simulation studies with this method, time is saved and different protocols can be defined. For example, there is no difficulty in the case of a design with correlated variables, as according to the variance formula for correlated variables, only it is needed to modify the random part and use the described algorithm.

The bisection method was used to find the final solution (the desired RD was reached after few iterations).

Different authors have proposed estimators for causal RD (6, 10-14). The results of the current study can be used to compare performance of these estimators.

For each hypothetical scenario, with the current method, all computations will be completed within almost 5 minutes with a system with a configuration of random access memory (RAM) 16.8 gigabytes (GB) and Core™ i7-4770 central processing unit (CPU) 3.40 GHZ, which is very helpful in large simulation studies. Additionally, an iterative method will not be necessary for estimating the coefficient of treatment.

Conclusion

To summarize, the proposed method in this

study is recommended over the current method due to less time consumption; this issue is important in studies with different scenarios.

Conflict of Interests

Authors have no conflict of interests.

Acknowledgments

The authors wish to acknowledge Dr. Sadaf Sepanlou for critical editing of the English grammar and syntax of the manuscript, and Dr. Mahboubeh Parsaeian and Dr. Leila Janani for their helpful comments.

References

1. Austin PC. A data-generation process for data with specified risk differences or numbers needed to treat. Commun Stat Simul Comput 2010; 39(3): 563-77.
2. 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.
3. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006; 60(7): 578-86.
4. Leemis LM, Shih LH, Reynertson K. Variate generation for accelerated life and proportional hazards models with time dependent covariates. Stat Probab Lett 1990; 10(4): 335-9.
5. Austin PC, Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. Commun Stat Simul Comput 2008; 37(6): 1039-51.
6. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med 2005; 24(11): 1713-23.
7. Lunn AD, Davies SJ. A note on generating correlated binary variables. Biometrika 1998; 85(2): 487-90.
8. Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. Stat Med 2012; 31(29): 3946-58.
9. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. Epidemiology 2015; 26(4): 466-72.
10. Balzer LB, Laan MJ. Estimating effects on rare outcomes: Knowledge is power [Online]. [cited 2013]; Available from: URL: https://biostats.bepress.com/ucbbiostat/paper310
11. Rosenblum M, van der Laan MJ. Simple Examples of Estimating Causal Effects Using Targeted Maximum Likelihood Estimation [Online]. [cited 2010]; Available from: URL: https://biostats.bepress.com/ucbbiostat/paper262
12. Bender R, Kuss O. Methods to calculate relative risks, risk differences, and numbers needed to treat from logistic regression. J Clin Epidemiol 2010; 63(1): 7-8.
13. Bender R, Blettner M. Calculating the "number needed to be exposed" with adjustment for confounding variables in epidemiological studies. J Clin Epidemiol 2002; 55(5): 525-30.
14. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 1998; 280(19): 1690-1.