Journal of the Korean Data & Information Science Society
2014, **25**(5), 1127–1135

http://dx.doi.org/10.7465/jkdi.2014.25.5.1127
한국데이터정보과학회지

Adjusting sampling bias in case-control genetic association studies[†]

Geum Chu Seo[1] · Taesung Park[2]

[1]Department of Statistics, Seoul National University
[2]Department of Statistics and Interdisciplinary Program in Bioinformatics, Seoul National University

Received 30 June 2014, revised 25 July 2014, accepted 4 August 2014

Abstract

Genome-wide association studies (GWAS) are designed to discover genetic variants such as single nucleotide polymorphisms (SNPs) that are associated with human complex traits. Although there is an increasing interest in the application of GWAS methodologies to population-based cohorts, many published GWAS have adopted a case-control design, which raise an issue related to a sampling bias of both case and control samples. Because of unequal selection probabilities between cases and controls, the samples are not representative of the population that they are purported to represent. Therefore, non-random sampling in case-control study can potentially lead to inconsistent and biased estimates of SNP-trait associations. In this paper, we proposed inverse-probability of sampling weights based on disease prevalence to eliminate a case-control sampling bias in estimation and testing for association between SNPs and quantitative traits. We apply the proposed method to a data from the Korea Association Resource project and show that the standard estimators applied to the weighted data yield unbiased estimates.

Keywords: Case-control design, genome-wide association studies, quantitative traits, sampling bias, SNPs.

1. Introduction

Genome-wide association studies (GWAS) have been successful in identifying genetic variants and their association with human quantitative traits. As described in Manolio *et al.* (2009), hundreds of GWAS have been conducted in the last few years and have identified over 1,000 trait-associated common single-nucleotide polymorphisms (SNPs), and the number continues to increase. A standard approach for GWAS analysis of quantitative traits has relied on linear regression analysis (i.e., classical least-squares estimation under the linear model). Exploring association between individual SNPs and human quantitative traits can

[†] This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP) (No. 2012R1A3A2026438).

[1] M.S. student, Department of Statistics, Seoul National University, Seoul 151-747, Korea.

[2] Corresponding author: Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea. E-mail: tspark@stats.snu.ac.kr

provide a crucial step in the discovery of disease susceptibility SNPs and valuable insights that are important for predicting disease risk and customizing treatment.

Many existing GWAS have adopted a case-control design, in which hundreds of thousands of SNPs are genotyped in a large number of cases (i.e., disease-affected individuals) and controls (i.e., disease-free individuals) in order to identify SNPs that are susceptible to diseases (Hunter *et al.*, 2007; Thomas *et al.*, 2008; Scott *et al.*, 2007). However, in this case-control designs, there are an issue that needs to be addressed to be able to analyze SNP-trait associations without bias. Because cases and controls are selected at different rates from their respective sub-populations, the case-control sample is not a random sample from the general population as described in Figure 1.1. In case-control studies, the proportion of cases is usually larger than the prevalence in the population, so bias may occur due to the disproportionate number of cases in the sample vs. the population.

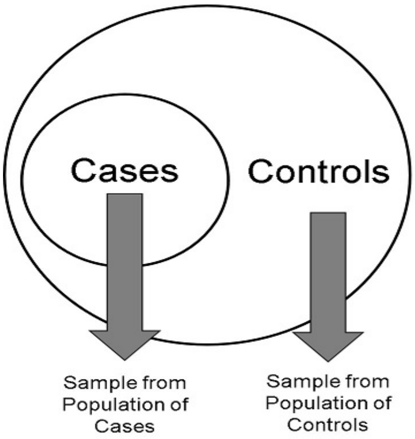
Cases
Controls
Sample from Population of Cases
Sample from Population of Controls

Figure 1.1 Case-control sampling design

As a result, the standard statistical analysis ignoring case-control sampling is likely to result in bias for assessing the effects of SNPs on quantitative traits. Although one may avoid the sampling bias by analyzing cases and controls separately or by including the case-control status as a covariate in a regression model, the associations between a genetic variant and a quantitative trait in the case and control groups can be quite different from the associations in the general population. These commonly-used analysis approaches can also provide biased estimates of the effect of SNPs on the quantitative traits.

In this paper, we propose a use of sampling weights in order to adjust for the sampling bias when estimating and testing for associations between individual SNPs and quantitative traits in case-control studies. Our method is based on inverse-probability of sampling weights with disease prevalence to account for case-control sampling bias. We apply the proposed method to eliminate the bias of the case-control sampling design, when we investigate SNP association tests with low-density lipoprotein trait by using a data from the Korea Association Resource (KARE) project (Cho *et al.*, 2009). We show that the sampling weighted regression provides unbiased estimates of SNP-trait associations.

2. Material and methods

The goal of the proposed method is to adjust for sampling bias in estimation and testing for SNP effects on the quantitative traits in case-control designs. First, we describe the proposed sampling weight method compared to VanderWeele and Vansteelandt (VanderWeele and Vansteelandt, 2010) weight method below, and then apply these weights to fit the linear regression model for quantitative traits.

2.1. VanderWeele and Vansteelandt weights

VanderWeele and Vansteelandt (2010) proposed the weight w_i which is specified to upweight the control individuals and downweight the affected individuals when the disease in the population is rare, as in

$$w_i = \begin{cases} \dfrac{\pi}{p_n} & \text{if } D_i = 1 \\ \dfrac{1-\pi}{1-p_n} & \text{if } D_i = 0 \end{cases} \tag{2.1}$$

where D_i is an indicator of an affected or control (1/0) individual, π is a disease prevalence in the population, and $p_n = n_1/n$ is the proportion of cases in the case-control study (i.e. the ratio of the number of cases in the study to the sum of the numbers of cases and controls in the study).

2.2. Inverse-probability of sampling weights based on disease prevalence

We consider the case-control study design when the proportions of cases and controls are not a random sample from the population. For a case-control study with a total of n subjects ($i = 1, ..., n$), let $S_{1,i}$ denote the selection status of individuals (1 = if sampled; 0 = otherwise) from population of cases and $S_{0,i}$ denote the selection status of individuals (1 = if sampled; 0 = otherwise) from population of controls. Let $p_{1,i}$ and $p_{0,i}$ denote the corresponding probabilities of being selected in cases and controls, respectively. Also, let x_i denote a vector of covariates such as age, area, and sex.

We use logistic regression models describing probability of selection in cases and controls separately as a function of predictor variable x_i. We also perform variable selection to find the best subset of predictors. The models can be written as

$$\text{logit}(p_{1,i}(x_i)) = \log\left(\frac{p_{1,i}(x_i)}{1-p_{1,i}(x_i)}\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi} \text{ for cases and} \tag{2.2}$$

$$\text{logit}(p_{0,i}(x_i)) = \log\left(\frac{p_{0,i}(x_i)}{1-p_{0,i}(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \text{ for controls.} \tag{2.3}$$

Equations (2.2) and (2.3) give the probabilities of selection

$$p_{1,i}(x_i) = \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{1+\exp(\alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})} \text{ for cases and} \tag{2.4}$$

$$p_{0,i}(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1+\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})} \text{ for controls.} \tag{2.5}$$

By using Equations (2.4) and (2.5), we can derive the sampling weights that are proportional to $1/p_{1,i}(x_i)$ and $1/p_{0,i}(x_i)$ respectively, because the weight of a sampled unit is the reciprocal of its probability of selection into the sample. The sampling weights reflecting the selection probabilities should produce representative probabilities of selecting case and control separately from the total population. Formally, the weights are computed by solving

$$\begin{aligned} w_{1,i}p_{1,i}(x_i) &= \tau_{1,i} \ (i = 1, \cdots, n_1) \\ w_{0,i}p_{0,i}(x_i) &= \tau_{0,i} \ (i = 1, \cdots, n_0) \end{aligned} \tag{2.6}$$

where n represents the sum of the total number of control individuals (n_0) and total number of affected individuals (n_1) sampled (that is, $n = n_0 + n_1$), and weights $w_{1,i}$ and $w_{0,i}$ are proportional to the inverse probability that individual i was sampled based on the logistic regression models in the case data set and the control data set. $\tau_{1,i}$ and $\tau_{0,i}$ are the probabilities of selecting case and control from the total population. If we define D_i and π as is in the section 2.1, then $\tau_{1,i}$ and $\tau_{0,i}$ can be written as

$$\begin{aligned} \tau_{1,i} = \cdots = \tau_{1,n_1} &= P(D_i = 1|Population) = \pi \\ \tau_{0,i} = \cdots = \tau_{0,n_0} &= P(D_i = 0|Population) = 1 - \pi. \end{aligned} \tag{2.7}$$

The sampling weights to adjust for case-control sampling bias are a function of disease prevalence, which is assumed to be known or estimated with external information. They also represent the inverse probability of unit selection into the sample. Therefore, the weight w_i can be specified as

$$w_i = \begin{cases} \dfrac{\pi}{p_{1,i}(x_i)} & \text{if } D_i = 1 \\ \dfrac{1-\pi}{p_{0,i}(x_i)} & \text{if } D_i = 0 \end{cases} \quad \text{with} \ \sum_{j=0}^{1}\sum_{i=1}^{n_j} w_{j,i} = n. \tag{2.8}$$

Both equations (2.1) and (2.8) specify the sampling weights to eliminate case-control sampling bias. The VanderWeele and Vansteelandt weights in equation (2.1) consider only two simple selection probabilities, which accounts for the proportion of cases and controls from the sample, respectively. However, the proposed sampling weight method is based on the selection probabilities of each sample, so this method is more specific to handle a non-random sampling issue for each individual than the method proposed by VanderWeele and Vansteelandt.

In addition, the proposed sampling weight method has a similarity with inverse probability weight (IPW) using the propensity score because both methods are designed to adjust sampling bias and use selection probabilities in order to derive weights. However, the proposed method can be differentiated from IPW approach using the propensity score in that the proposed method computes selection probabilities for cases and controls separately using two logistic regression models in Equations (2.2) and (2.3). The modeling for selection probabilities obtained from the proposed method is due to sampling design; cases and controls are selected at different rates from their respective sub-population, so they are not random samples from the general population. IPW using the propensity score weights each treated individual with the inverse of its propensity score and all untreated individuals are weighted by the inverse of one minus its propensity score (Hernan and Robins, 2006). Therefore, only one logistic regression is needed for IPW using the propensity score since the fit predicts both the probability p of being in treatment group and that of "not treatment group", $= 1 - p$.

2.3. Correction for sampling bias in estimating and testing association between genetic variants and quantitative traits

Suppose that quantitative trait y_i and a vector of covariates x_i are observed for the i^{th} of n individuals. A standard approach to test SNP-trait associations is to fit a linear regression as

$$y_i = \alpha_0 + \alpha_1 x_i + \beta SNP_i + \epsilon_i. \tag{2.9}$$

When we fit the linear regression model (2.9) for quantitative traits using case-control data, the case-control study design cannot be ignored. In the case-control study designs, we apply the proposed weights and the weights from VanderWeele and Vansteelandt to the equation (2.9) in order to adjust for sampling bias for association between genetic variants and quantitative traits. When the weights are used, the case-control study design can effectively be ignored. If we fit a linear regression of SNP on quantitative trait using the case-control data but weighting each case and control by the equations (2.1) and (2.8), then the coefficients obtained in this weighted regression will give an unbiased estimator of β obtained in a linear regression of SNPs on quantitative traits using data from a cohort study of the same population.

2.4. Real data application

This study used the data from the Korea Association Resource (KARE) project, which has been described elsewhere (Cho *et al.*, 2009). KARE study undertook a large scale genome-wide association studies (GWAS) for human complex quantitative traits among 10,038 participants aged 40 to 69. About 10,000 subjects from KARE study cohorts were genotyped with Affymetrix Genome-Wide Human SNP array 5.0. The individuals were recruited from two prospective population-based studies as part of the Korean Genome Epidemiology Study project, the rural Ansung (n = 5,018) and urban Ansan (n = 5,020) cohorts, in Gyeonggi Province, South Korea. Both cohorts were designed to allow longitudinal prospective study and adopted the same investigational strategy. The standard quality control procedures (Cho *et al.*, 2009) were adopted and we included GWA genotypes from 8,842 individuals in the association analysis.

All the individuals were measured for a range of quantitative traits related to obesity, blood condition, and lipids. We examined low-density lipoprotein trait related to type 2 diabetes (T2D).

Among a total of 8,842 KARE study participants, 526 and 560 subjects are sampled as T2D cases and controls according to the following criteria. The inclusion criteria of T2D case subjects were as follows: (1) treatment of T2D, (2) fasting plasma glucose $\geq$ 7 mmol/L or plasma glucose 2-h after ingestion of 75 gm oral glucose load $\geq$ 11.1 mmol/L and (3) age of disease onset $\geq$ 40 years. The inclusion criteria of nondiabetic control subjects were as follows: (1) no history of diabetes and (2) fasting plasma glucose <5.6 mmol/L and plasma glucose 2-h after ingestion of 75gm oral glucose load <7.8 mmol/L at both baseline and follow up studies. Only these selected samples (T2D consortium sample) were used for generating the whole genome sequencing data due to high cost of sequencing.

3. Results

To demonstrate our method, we applied the proposed weights to test for association between SNPs and low-density lipoprotein (LDL) trait in the T2D consortium sample of 526 T2D cases and 560 controls. Since the 1086 T2D consortium sample was not randomly selected from 8,842 subjects from KARE cohort data, this sample can be regard as a biased sample induced by the case-control design.

We fit a standard linear regression from equation (2.9) to KARE data (population) and T2D consortium sample, and age, area, and sex information is used as covariates in the analysis. In order to eliminate the bias from the case-control sampling design, we fit a weighted linear regression to T2D consortium sample by using the proposed weights in equation (2.8) and the weights from VanderWeele and Vansteelandt in equation (2.1). We then compare these regression results in the following settings: significant SNPs vs non-significant SNPs from population-based genome-wide association studies (see Table 3.1 and Table 3.2)

We selected the top six SNPs from the population-based association studies where p-value is less than 1E-08. According to Table 3.1, these significant SNPs, in the population-based association studies, are found to be insignificant in the unadjusted association studies with the T2D consortium sample. These inconsistent results are caused by sampling bias from the case-control study design. After the weight adjustment, the adjusted p-values for most SNPs become closer to those obtained from population-based association studies. This illustrates that the weight adjustment methods successfully adjust the p-values from the T2D consortium sample.

Table 3.1 SNPs significantly associated with LDL trait in population-based genome-wide association studies

		Population		Unadjusted		Adjusted with the e proposed weight		Adjusted with VanderWeele and Vansteelandt weight	
CHR	SNP	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
1	rs599839	-0.0069	1.69E-12	-0.0045	0.1233	-0.0036	0.2270	-0.0031	0.2931
5	rs10942739	0.0031	1.75E-09	-0.0001	0.9511	0.0021	0.1822	0.0018	0.2499
5	rs12654264	-0.0029	1.87E-09	-0.0021	0.1479	-0.0042	0.0051	-0.0041	0.0063
5	rs4045166	0.0031	2.38E-09	-0.0002	0.8930	0.0023	0.1480	0.0019	0.2244
5	rs3846663	-0.0029	3.78E-09	-0.0021	0.1454	-0.0042	0.0051	-0.0041	0.0063
10	rs12242220	0.0128	7.08E-09	0.0187	0.0518	0.0234	0.0010	0.0225	0.0018

Abbreviations: CHR, a chromosome; Est., a regression coefficient

In addition, ten non-significant SNPs (p-value>0.05) from the population-based association studies were randomly chosen from chromosomes 1 to 10 and are summarized in Table 3.2. This table shows that non-significant SNPs in the population-based association studies found to be significant in the association studies of the T2D consortium sample without weight adjustment. These inconsistent results are due to sampling bias in case-control study. However, these SNPs turned out to be not significant after eliminating the sampling bias by the sampling weight methods. This illustrates that the adjusted weighting methods properly adjust the p-values from the T2D consortium sample.

In summary, Tables 3.1 and 3.2 show that the proposed sampling weights would adequately adjust for the sampling bias induced by the case-control sampling design.

Table 3.2 SNPs not significantly associated with LDL trait in population-based genome-wide association studies

		Population		Unadjusted		Adjusted with the proposed weight		Adjusted with VanderWeele and Vansteelandt weight	
CHR	SNP	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
1	rs521179	-0.0020	0.1062	-0.0101	0.0134	-0.0075	0.0998	-0.0086	0.0556
2	rs407427	0.0011	0.0690	0.0045	0.0138	0.0036	0.0606	0.0045	0.0167
3	rs17072261	0.0011	0.0898	0.0045	0.0167	0.0033	0.0724	0.0039	0.0367
4	rs2691389	-0.0011	0.0785	-0.0045	0.0182	-0.0033	0.1047	-0.0044	0.0267
5	rs11742326	-0.0008	0.1210	-0.0030	0.0398	-0.0019	0.2109	-0.0022	0.1400
6	rs9344742	0.0013	0.2191	0.0074	0.0240	0.0038	0.2691	0.0049	0.1557
7	rs1008454	-0.0004	0.3695	-0.0030	0.0402	-0.0015	0.3044	-0.0024	0.1094
8	rs1499425	0.0005	0.2777	0.0040	0.0063	0.0014	0.3559	0.0022	0.1397
9	rs4237110	-0.0003	0.5795	-0.0037	0.0206	-0.0009	0.5733	-0.0016	0.3138
10	rs11239821	-0.0015	0.4710	-0.0221	0.0046	-0.0033	0.7044	-0.0075	0.3609

Abbreviations: CHR, a chromosome; Est., a regression coefficient

In order to see how p-values from the weight adjustment methods are close to those of the population-based association studies, the p-values from T2D consortium sample were compared with those from population-based association study. As shown in Figure 3.1, 29 SNPs were plotted whose population-based p-values are less than 1E-05. Numbers from 1 to 29 represent the corresponding p-value of the selected SNP. X-axis represents the p-values from the population-based association study and Y-axis does those from T2D consortium data analyses. Each plotted number represents the result of T2D consortium sample with three colors: black color for unadjusted analysis, red color for adjusted analysis using the proposed sampling weight method, and blue color for adjusted analysis using the weight from VanderWeele and Vansteelandt. The regression slopes of these three colors are -0.06, 0.06, and 0.03 for unadjusted and weight adjusted methods, respectively. The closer to 1 the slope is, the better the method is. The p-values of the weight adjusted methods become closer to p-values from population-based association study than those from unadjusted method.

We examined the Pearson's correlation coefficient for p-values between population and T2D consortium sample from SNP-LDL trait association study; the results are summarized in Table 3.3 showing that p-values between population and adjusted sample with the sampling weights are more highly correlated than those between population and biased sample.

Table 3.3 Pearson's correlation coefficient with p-values between population and sample from SNP-LDL trait association study

		Pearson's correlation coefficient with p-values		
P-value Threshold	SNP #	Unadjusted	Adjusted with the proposed weight	Adjusted with VanderWeele and Vansteelandt weight
1E-05	29	-0.0700	0.4882	0.4333
1E-04	94	0.0221	0.2119	0.1864
1E-03	545	0.0616	0.0822	0.0821

In addition, we compared test statistics described in Figure 3.2 under the similar settings with the p-value comparison, and the regression slopes of test statistics were 0.20, 0.32, and 0.30 for unadjusted and weight adjusted methods, respectively. We obtained similar result as we observed in Figure 3.1. That is, the test statistics of the weight adjusted methods for T2D consortium sample become closer to those from the population-based association study

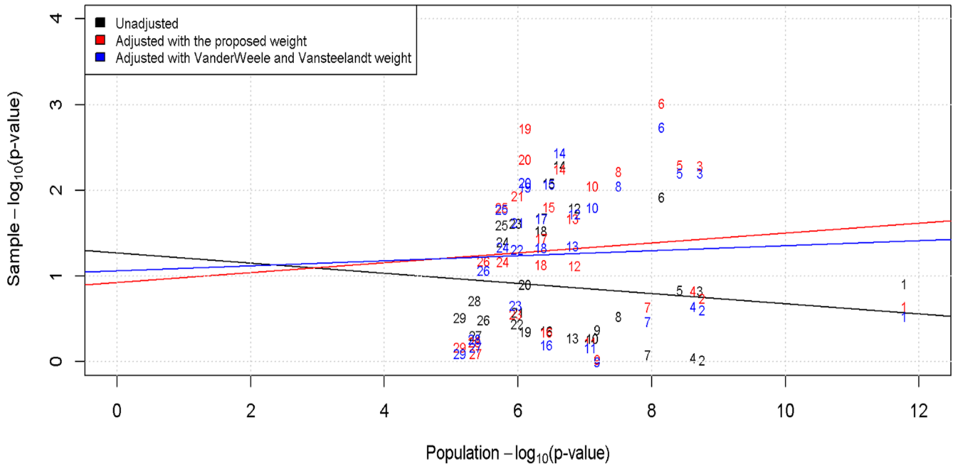

Figure 3.1 Comparison of p-values (population-based p-value threshold < 1E-05) using unadjusted samples (black-colored), adjusted samples with the proposed weight (red-colored), and adjusted samples with VanderWeele and Vansteelandt weight (blue-colored)

than those from unadjusted method.

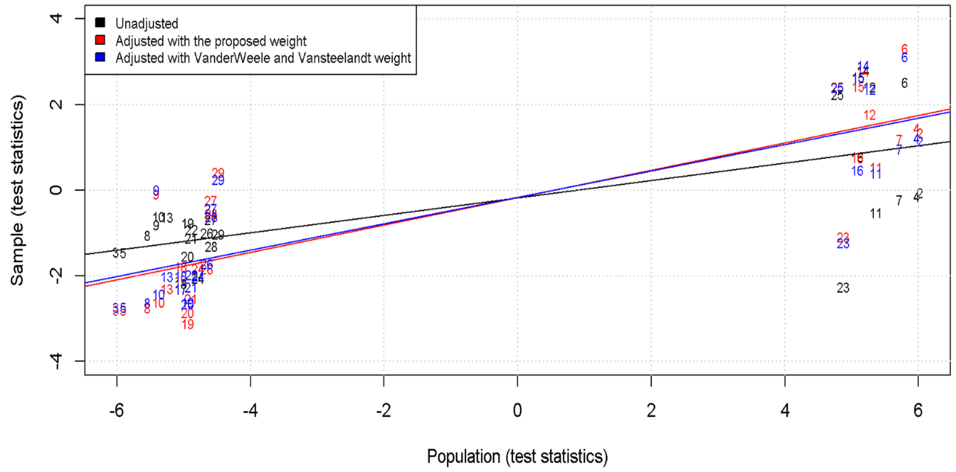

Figure 3.2 Comparison of test statistics (population-based p-value threshold < 1E-05) using unadjusted samples (black-colored), adjusted samples with the proposed weight (red-colored), and adjusted samples with VanderWeele and Vansteelandt weight (blue-colored)

We also examined the Pearson's correlation coefficient for test statistics described in Table 3.4 and obtained similar results as we observed in the p-value analysis; test statistics between population and adjusted sample using the proposed weight are highly correlated than those between population and biased sample.

Therefore, our case-control weighting scheme seems to successfully estimate the regression coefficients and p-values even with poor or biased samples of data.

Table 3.4 Pearson's correlation coefficient with test statistics between population and sample from SNP-LDL trait association study

		Pearson's correlation coefficient with test statistics		
P-value Threshold	SNP #	Unadjusted	Adjusted with the proposed weight	Adjusted with VanderWeele and Vansteelandt weight
1E-05	29	0.6952	0.8400	0.8311
1E-04	94	0.7099	0.7164	0.7273
1E-03	545	0.6032	0.6136	0.6504

4. Discussion

The bias induced by the case-control design might be a concern when the goal of the study is characterizing the relationship between SNP and a quantitative trait. In this study, we considered using the inverse-probability of sampling weights based on disease prevalence to correct sampling bias when estimating and testing for SNP-trait associations in a case-control study design. We have shown that when sampling fractions are known, the weighted regression provides much less biased estimators of measures of association between the marker and trait in the case-control study.

Our real data analysis shows that the analysis using the weight from VanderWeele and Vansteelandt seems to work well in adjusting for sampling bias induced by case-control study. The weights in equation (2.1) are only based on two simple selection probabilities, which accounts for the proportion of cases and controls from the sample, respectively. On the other hand, our proposed sampling weight method considers the selection probability of each individual. Therefore, our method is more specific to handle a non-random sampling issue for each individual, and is a more flexible approach to adjust for sampling bias than the method suggested by VanderWeele and Vansteelandt.

References

Cho, Y. S., Go M. J., Kim Y. J., Heo J. Y., Oh J. H., Ban, H., Yoon D., Lee, M. H., *et al.* (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genetics*, **41**, 527-534.

Hernan, M. A. and Robins J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, **60**, 578-586.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., *et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, **39**, 870-874.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature Genetics*, **461**, 747-753.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., *et al.* (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341-1345.

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., *et al.* (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, **40**, 310-315.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, **172**, 1339-1348.