Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Abarin, J Biomet Biostat 2013, 4:4
DOI: 10.4172/2155-6180.1000173

Research Article

Open Access

Gene-environment Interaction Studies with Measurement Error Application in the Complex Diseases in the Newfoundland Population: Environment and Genetics Study

Taraneh Abarin*

Department of Mathematics and Statistics, Memorial University, Canada

Abstract

Newfoundland and Labrador (NL) has had the highest percentage of overweight/obese residents in Canada since 2007. This complex trait is determined by multiple genetic and environmental factors that interact with one another. The existing studies examine such factors under the assumption that they are measured accurately. However, error- prone environmental and genetic factors are unavoidable. The impact of ignoring these errors varies from bias to false results in detecting associations. Motivated by CODING study, we present methodologies to estimate model parameters, while accounting for measurement error and misclassification. We applied bias-corrected methods for three separate studies: candidate-gene association study and two gene-environment interaction models, where both environmental and genetic factors are subject to error. Our results based on simulation studies show that the proposed methodologies perform quite satisfactory.

Keywords: Bias-corrected; Gene-environment interaction; Measurement error; Genotyping error; Misclassification

Abbreviations: CODING: Complex Diseases in the Newfoundland Population Environment and Genetics; BC: Bias-corrected; ME: Measurement Error; GEI: Gene-environment Interaction; PA: Physical Activity; PTF: Percent Trunk Fat; FTO: Fat Mass and Obesity

Introduction

Obesity is a major health issue in Canada. Newfoundland and Labrador (NL) has had the highest percentage of overweight/obese residents in Canada since 2007, and had risen by nearly 7% to 69.3% in 2011 (Statistics Canada). Obesity is determined by multiple genetic and environmental factors that interact with one another in complicated ways. The existing studies examine such factors under the assumption that they are measured accurately [1-4]. However, unobserved or error-prone environmental factors, and/or misclassification in genotyping are unavoidable. In reality, both genetics and environmental factors are likely measured with errors. It is now well-known that measurement, and/or classification errors can influence the results of a study. The impact of ignoring these errors varies from bias and large variability in estimators to low power or even false-negative results in detecting genetic associations [5-7]. In fact, in the presence of measurement error and misclassification, detecting the interaction terms is more challenging than either the genetic or the environmental factors [8]. Motivated by an ongoing, large scale nutrigenomics (CODING) study of Newfoundland adults' population, we present methodologies to estimate model parameters, while accounting for measurement error and misclassification. We applied bias-corrected methods to three separate studies: candidate gene association study and two gene-environment interaction models, where both environmental and genetic factors are subject to error. This paper is organized as follows. In Section 2, we introduce the three models, and present bias-corrected estimators. We investigate the finite sample performances of the proposed estimators in comparison with the naive estimators, using some simulation studies, in Section 3. The estimation approaches are also illustrated in this section, with the analysis of the CODING data.

Materials and Methods

Model I: Candidate gene association study

Motivated by the CODING study of Newfoundland population, we present the methodologies to estimate the model parameters, for three separate studies: candidate gene association study, and two different GEI models. In all these three models, we assume that the response is measured accurately.

In this section, we consider a simple linear regression model for typical candidate-gene association studies. The model can be written as

$$Y_i = \beta_0 + \beta_1 G_i + \in_i, \quad i = 1, \cdots, n, \tag{1}$$

Where $Y_i \in \mathbb{R}$ is the response for the i^{th} individual, β_0 and β_1 $\in \mathbb{R}$ are unknown parameters, and G is a binary variable, coded for a candidate gene with dominant effect. One can write model (1) in matrix format as

$$Y = \tilde{X}\beta + \in, \tag{2}$$

where $Y_{n\times 1}$ is the vector of response, $\in_{n\times 1}$ is the vector of model error terms with mean zero and variance $\sigma^2_{\in}, \beta = (\beta_0, \beta_1)'$ the vector of parameters, and $\tilde{X} = [1, G]$ is the $n \times p$ design matrix.

Moreover, binary variable G with probability of success π is not observable, and instead a binary variable g is observed with classification error. We denote sensitivity or probability of correctly

***Corresponding author:** Taraneh Abarin, Department of Mathematics and Statistics, Memorial University, St. John's, NL, Canada, Tel: (709) 864-8733; Fax: (709) 864-3010; E-mail: tabarin@mun.ca

Received July 29, 2013; **Accepted** August 26, 2013; **Published** August 30, 2013

Citation: Abarin T (2013) Gene-environment Interaction Studies with Measurement Error Application in the Complex Diseases in the Newfoundland Population: Environment and Genetics Study. J Biomet Biostat 4: 173. doi:10.4172/2155-6180.1000173

classifying success in G, with $\theta_{11} = P(g=1|G=0)$. Therefore, $1-\theta_{11}$ is the probability of false-negative. Similarly, specificity or probability of correctly classifying failure in G, is defined as $\theta_{00} = P(g=0|G=0)$. Therefore, $1-\theta_{00}$ is the probability of false-positive. The probability of success for g, is determined as follows.

$$p(g=1) = p(g=1|G=1)p(G=1) + p(g=1|G=0)p(G=0) = \theta_{11}\pi + (1-\theta_{00})(1-\pi).$$

In order to obtain an unbiased estimate for π based on the observed variable, one must correct the bias in g. In fact, with a simple algebra an unbiased estimate for π based on g is

$$g_{bc} = \frac{g-1+\theta_{00}}{\theta_{00}+\theta_{11}-1}.$$

The naive Least Squared estimator of β in model (1) that ignores the misclassification in g, is $\hat{\beta}_{naive} = (X'X)^{-1}X'Y$, where $X = [1, g]$.

Rewriting model (1) based on the observed variable g as follows; it is easy to see that $\hat{\beta}_{0_{naive}}$ is an unbiased estimator.

$$\begin{aligned} E(Y|X) &= E_{X|\tilde{X}}(E(Y|\tilde{X}, X)|\tilde{X}) \\ &= E_{X|\tilde{X}}(E(Y|\tilde{X})\tilde{\mathrm{X}} \\ &= E_{X|\tilde{X}}(\beta_0 + \beta_1 G|\tilde{X}) \\ &= \beta_0 + \beta_1 E(G|g) \\ &= \beta_0 + \beta_1 P(G=1|g) \end{aligned}$$

In the second equation, X or actually g is assumed to be surrogate, which means that it does not provide any extra information about the distribution of Y given what is already provided by G.

From the above equations, it can be seen that $\hat{\beta}_{1_{naive}}$ is biased. It is known that this bias is attenuated with large sample size [5,6]. Furthermore, when π is not very small, the naïve estimator is sensitive to sensitivity, in the sense that the smaller θ_{11}, the worse the naïve estimator [8].

Modifying the methodology suggested by Buonaccorsi [9] for the linear model with an intercept, the matrix of classification probabilities is defined as

$$\Theta = \begin{bmatrix} \theta_{11} & 1-\theta_{00} \\ 1-\theta_{11} & \theta_{00} \end{bmatrix}.$$

Using the same notation as [9], we have the mean responses for both genotyping groups as $\mu_1 = \beta_0 + \beta_1$ and $\mu_2 = \beta_0$.

Bounacccorsi (9) proposed a bias-corrected estimator for $\mu = (\mu_1, \mu_2)$ as $\hat{\mu} = (\Theta D_x)^{-1} gg$, where $gg = (g, 1-g)$, $D_x = diag(\hat{n})$, $\hat{n} = \Theta^{-1} n_w$, and n_w is defined as

$$n_w = \begin{bmatrix} n_{w1} \\ n_{w2} \end{bmatrix},$$

With n_{w1} to be the number of successes in the sample and $n_{w2} = n - n_{w1}$ number of failures in the sample. Returning back the estimates based on β, we have

$$\hat{\beta}_{bc} = \begin{bmatrix} \hat{\mu}_2 \\ \hat{\mu}_1 - \hat{\mu}_2 \end{bmatrix}$$

This method can be easily extended to any candidate gene with an additive effect. We should mention in here that genotyping error is usually estimated in two different ways. There are either two different methods of genotyping compared, or genotyping using one system is repeated more than once. The later is less expensive.

Model II: Gene-environment interaction I

Now, we consider the first GEI model as

$$Y_i = \beta_0 + \beta_1 G + \beta_2 W + \beta_3 G * W + \beta_4 A + \varepsilon. \tag{3}$$

In this model, an environmental factor $W \in \mathbb{R}$ is unobservable. Instead, one observes Z subject to certain measurement error. The measurement error (classic) model may be expressed as

$$Z = W + U, \tag{4}$$

where U is an unobservable measurement error variable, independent from W, with mean zero and variance, say σ_u^2. We also observe g (instead of G) with error. In model (3), there is another environmental factor (A), which is assumed to be measured without error. The interaction term $G*W$ in the model, is between two error-prone variables. We are interested in estimating $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$.

Defining X to be the designed matrix based on the observed variables $[1, g, Z, gZ, A]$, the naive estimator that ignores both ME and misclassification in the variables, can be expressed as

$$\hat{\beta}_{naive} = (X'X)^{-1}X'Y$$

$$= \begin{bmatrix} n & \sum g & \sum Z & \sum gZ & \sum A \\ \sum g & \sum g^2 & \sum gZ & \sum g^2 Z & \sum gA \\ \sum Z & \sum gZ & \sum Z^2 & \sum gZ^2 & \sum ZA \\ \sum gZ & \sum g^2 Z & \sum gZ^2 & \sum g^2 Z^2 & \sum gZA \\ \sum A & \sum gA & \sum ZA & \sum gZA & \sum A^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum gY \\ \sum ZY \\ \sum gZY \\ \sum AY \end{bmatrix}, \tag{5}$$

where the sums in the matrices are over the number of observations.

The methodology suggested by Buonaccorsi [9] to correct the bias caused by misclassification, cannot be applied to this model. Since both sensitivity and specificity are large, the bias caused by this error is small [8]. However, the bias caused by U cannot be ignored. In fact, the larger the variability of U, the worse the naive estimator.

Since

$$E(\sum_i^n Z_i^2 | W) = \sum_i^n W_i^2 + \sigma_u^2, \tag{6}$$

$\sum Z^2$, $\sum gZ^2$, $\sum g^2 Z^2$ in equation 5 need to be corrected for bias. However, bias-correcting these terms requires σ_u^2 to be estimated. Generally, estimating σ_u^2 requires extra information, such as internal or external validation data [5,6]. The BC estimator of β, therefore, can be expressed as follows.

$$\hat{\beta}_{bc} = \begin{bmatrix} n & \sum g & \sum Z & \sum gZ & \sum A \\ \sum g & \sum g^2 & \sum gZ & \sum g^2 Z & \sum gA \\ \sum Z & \sum gZ & c_{33} & c_{34} & \sum ZA \\ \sum gZ & \sum g^2 Z & c_{34} & c_{34} & \sum gZA \\ \sum A & \sum gA & \sum ZA & \sum gZA & \sum A^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum gY \\ \sum ZY \\ \sum gZY \\ \sum AY \end{bmatrix},$$

where $c_{33} = \sum Z^2 - n\hat{\sigma}_u^2$, and $c_{34} = \sum gZ^2 - \sum g\hat{\sigma}_u^2$. Since g is binary, $E(\sum gZ^2) = E(\sum g^2 Z^2)$.

Moreover, since $E(U)=0$, there is no need for correcting the other terms in the naive estimator.

Model III: Gene-environment interaction II

Now, we consider the second GEI model as

$$Y_i = \beta_0 + \beta_1 G + \beta_2 A + \beta_3 G * A + \beta_4 W + \varepsilon. \qquad (7)$$

In this model again, both W and G are unobservable. However, in here, the interaction is between the misclassified variable and the accurately measured environmental factor. We are interested in estimating $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$.

Defining X to be the designed matrix based on the observed variables $[1, g, A, gA, Z]$, the naive estimator can be expressed as

$$\hat{\beta}_{naive} = (X'X)^{-1}X'Y$$

$$= \begin{bmatrix} n & \sum g & \sum A & \sum gA & \sum Z \\ \sum g & \sum g^2 & \sum gA & \sum g^2 A & \sum gZ \\ \sum A & \sum gA & \sum A^2 & \sum gA^2 & \sum AZ \\ \sum gA & \sum g^2 A & \sum gA^2 & \sum g^2 A^2 & \sum gZA \\ \sum Z & \sum gZ & \sum AZ & \sum gZA & \sum Z^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum gY \\ \sum AY \\ \sum gZY \\ \sum ZY \end{bmatrix} \qquad (8)$$

In here, only $\sum Z^2$ needs to be corrected for bias. The BC estimator of β, therefore, can be expressed as follows.

$$\hat{\beta}_{bc} = \begin{bmatrix} n & \sum g & \sum A & \sum gA & \sum Z \\ \sum g & \sum g^2 & \sum gA & \sum g^2 A & \sum gZ \\ \sum A & \sum gA & \sum A^2 & \sum gA^2 & \sum AZ \\ \sum gA & \sum g^2 A & \sum gA^2 & \sum g^2 A^2 & \sum gZA \\ \sum Z & \sum gZ & \sum AZ & \sum gZA & c_{33} \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum gY \\ \sum AY \\ \sum gZY \\ \sum ZY \end{bmatrix}$$

$$c_{33} = \sum Z^2 - n\hat{\sigma}_u^2.$$

Covariance matrices

Since the naive estimator does not consider Z or g as random variables, its covariance matrix can be easily written as

$$V(\hat{\beta}_{naive}) = \sigma_{\epsilon}^2 (X'X)^{-1}$$

The covariance matrix of the bias corrected estimator, however, is conditional on both g and Z, as follows

$$V(\hat{\beta}_{bc} \,|\, Z, g)) = \sigma_{\epsilon}^2 (X'X)_{bc}^{-1} X'X (X'X)_{bc}^{-1} \qquad (9)$$

Results

Simulation studies

To examine the finite-sample performance of the bias-corrected approaches for estimating the regression parameters, we carried out some simulation studies. For each model, we present the simulation set ups and the results, separately.

Model I: Candidate gene association study: For this model, we considered n=500 observations. The regression coefficients were $\beta = (2, 0.1)'$, and the model error variance was set to be $\sigma_{\epsilon}^2 = 2$. The response was generated 1,000 times, by using model (1). Both sensitivity and specificity were 0.95. We compared three, namely True (based on G), Naive (based on g), and BC estimation approaches.

Figure 1 exhibits the magnitude of biases produced by all the three approaches. From the figure we can clearly see that among the three estimators, True and BC estimators are performing well. It is also noted in here that, since both sensitivity and specificity were relatively large, the impact of misclassification on the estimators, is relatively small.

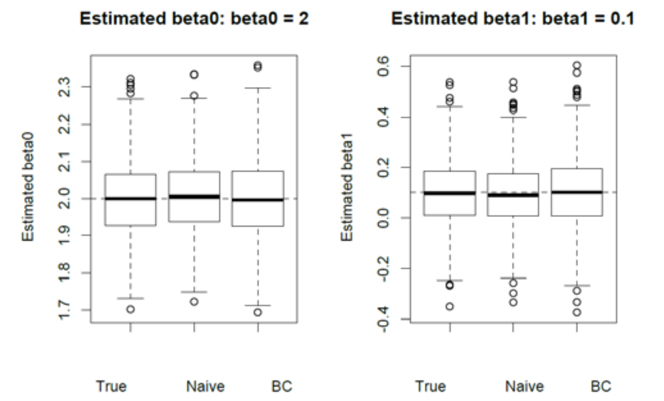

Figure 1: Box plots of True, Naïve and BC estimators for $Y = \beta_0 + \beta_1 G + \in$ with misclassified.

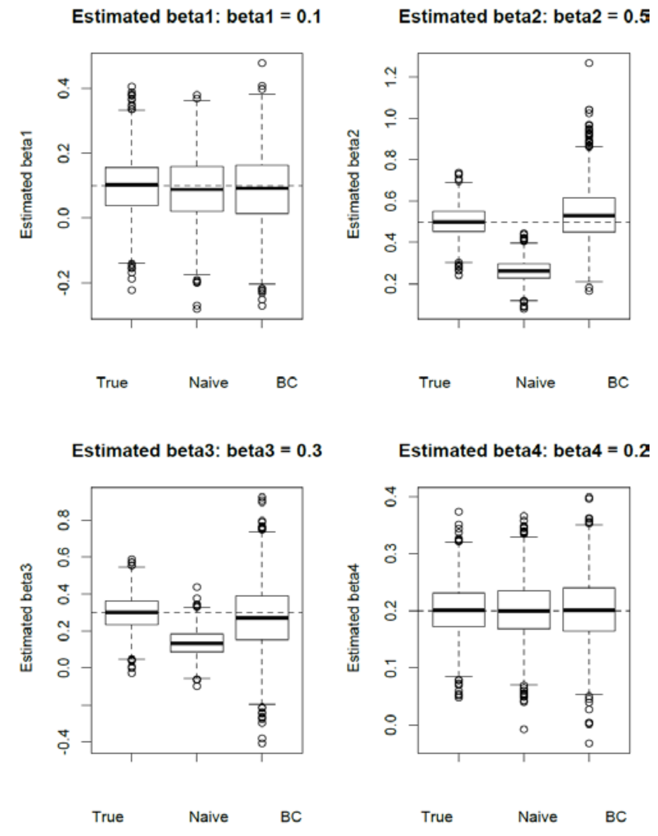

Figure 2: Box plots of True, Naïve, and BC estimators for $Y = \beta_0 + \beta_1 G + \beta_2 W + \beta_3 G * W + \beta_4 A + \in$ with error-prone *W* and *G*.

Model II: Gene-environment interaction I: For this model, we considered n=500 observations. The regression coefficients were $(2, 0.1, 0.5, 0.3, 0.2)'$, and the model error variance was set to be $\sigma_{\epsilon}^2 = 1$. The response was generated 1,000 times, by using model (3). Both sensitivity and specificity were 0.95. Environmental factor W and A, for simplicity, were generated from a standard normal distribution. The error-prone variable *Z* was generated from model *Z*=*W*+*U*, where *U* is independent of *W* and has normal distribution with mean zero and variance $\sigma_u^2 = 1.$ Here again, we compared the three approaches: True (based on *G* and *W*), Naive (based on *g* and *Z*), and BC estimation approach.

Figure 2 shows the magnitude of biases produced by the three approaches. From the figure we can see again that True and BC estimators are performing well. The naive use of *W* as *Z* causes remarkable biases in the estimators of β_2 and the coefficient of the interaction term β_3. It is also noted that, since the misclassification rates are low, the impact of misclassifications on the estimators are negligible. Moreover, since the naive estimate of β_0 is unbiased, the box plot for this parameter is omitted.

Model III: Gene-environment interaction II: For this model, we again considered n=500 observations. The regression coefficients were $\beta = (2, 0.1, 0.5, 0.3, 0.2)'$, and the model error variance was set to be $\sigma_\epsilon^2 = 1.$ The response was generated 1,000 times, by using model (7). Both sensitivity and specificity were 0.95. Environmental factor *W* and *A* were generated from a standard normal distribution. The error-prone variable *Z* was generated from model *Z*=*W*+*U*, where *U* is independent of *W* and has normal distribution with mean zero and variance $\sigma_u^2 = 1.$ Here again, we compared True, Naive and BC estimation approaches.

Figure 3 shows the magnitude of biases produced by the three approaches. From the figure we can see again that True and BC estimators are performing well. The naive use of *W* as *Z* causes remarkable bias in the estimator of β_4, the coefficient of *W*. It is also noted that since the misclassification rates were low, the impact of misclassifications on the estimators were negligible. Moreover, since the naive estimate of β_0 is unbiased, the box plot for this parameter is omitted.

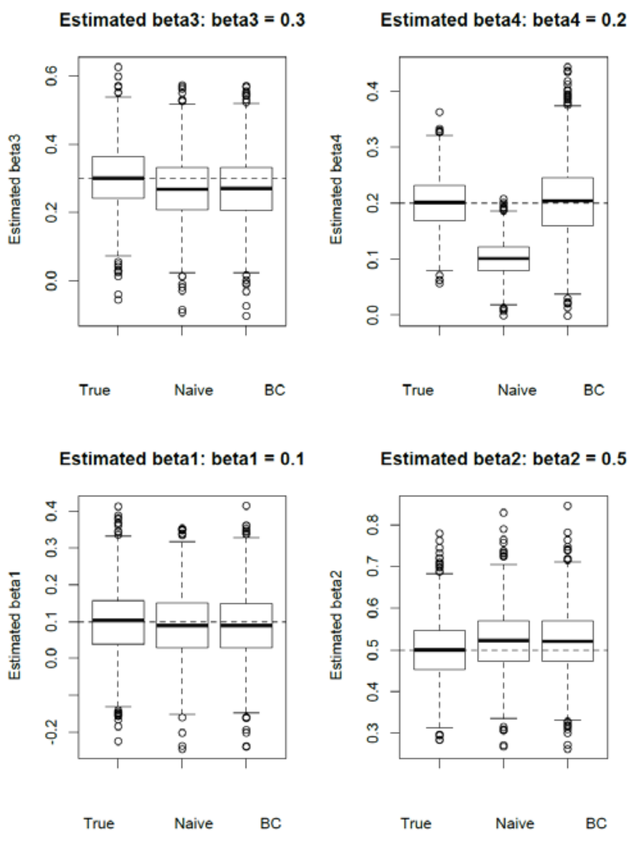

Figure 3: Box-plots of true, naïve and BC estimators for $Y = \beta_0 + \beta_1 G + \beta_2 A + \beta_3 G * A + \beta_4 W + \in$ with error-prone *W* and *G*.

Application: CODING study

Complex Diseases in the Newfoundland Population: Environment and Genetics (CODING) is an ongoing, large scale nutrigenomics study of Newfoundland population, in which 2256 individuals from the Newfoundland population were recruited. Variables considered were PTF measured by dual X-ray absoprtiometry as response, rs9939609 single-nucleotide polymorphisms of the FTO gene, genotyped using the high-throughput MassARRAY R platform (Sequenom Inc, San Diego, CA, USA), and PA measured by the Ability of the Atherosclerosis Risk in Communities (ARIC) Baecke et al. [10], questionaire as covariates. Subjects were stratified by gender for analysis. Gene-candidate association, gene-physical activity interaction, and gene-age interaction were studied. PTF was assumed to be measured with no error. Age was also assumed to be measured accurately. To avoid the colinearty between the variables, age was centred around its mean.

Combination of Sports and Leisure Time Index was selected for the analysis of PA, which was assumed to be measured with error. FTO were coded as *G*=1 and *G*=0, for "A" allele with dominant effect. Genotyping error was estimated to be 5%. The purpose of our study was to estimate the coefficients of the three models, accounting for measurement error and genotyping error.

Since there was no extra information available to estimate σ_u^2, we performed a sensitivity analysis. It should be mentioned in here that for the ME model (4), σ_Z^2 is always larger than σ_u^2. In the CODING data, the sample variance for the observed PA was 1.3. Therefore, two arbitrary values of 0.1 and 0.5 were chosen as representatives for relatively small and relatively large values for σ_u^2, respectively. Evidently, the larger the value for σ_u^2, the worse the naive estimates of the parameters! Naive (based on observed genotyping and PA) and BC approach (bias-corrected for errors) estimates were calculated for each model with their corresponding standard errors. $\hat{\sigma}_\epsilon^2$ was calculated using the naive least squared estimators of each model.

Tables 1 and 2 show the results for males and females, separately.

	Naive	SE	BC $(\sigma_u^2 = 0.1)$	SE	BC $(\sigma_u^2 = 0.5)$	SE
Model I						
β_0	28.18	0.68	27.92	0.50	27.92	0.50
β_1	2.19	0.83	2.50	0.75	2.50	0.75
Model II						
β_0	33.85	3.06	34.70	3.20	39.27	3.93
β_1	0.70	3.38	0.75	3.55	1.25	4.46
β_2	-2.17	0.41	-2.28	0.43	-2.90	0.54
β_3	0.15	0.52	0.14	0.54	0.05	0.69
β_4	0.21	0.03	0.21	0.02	0.19	0.03
Model III						
β_0	33.35	2.67	34.31	2.76	39.59	3.29
β_1	1.49	2.19	1.39	2.18	0.89	2.19
β_2	0.21	0.04	0.20	0.04	0.18	0.04
β_3	0.004	0.05	0.006	0.05	0.01	0.05
β_4	-2.08	0.26	-2.20	0.27	-2.88	0.36

Model I: $PTF = \beta_0 + \beta_1 G + \in,$
Model II: $PTF = \beta_0 + \beta_1 G + \beta_2 PA + \beta_3 PA * G + \beta_4 Age + \in,$
Model III: $PTF = \beta_0 + \beta_1 G + \beta_2 Age + \beta_3 Age * G + \beta_4 PA + \in$
Table 1: Estimates of model coefficients and the standard errors of naive and BC approach for CODING study–Males.

	Naive	SE	BC $(\sigma_u^2 = 0.1)$	SE	BC $(\sigma_u^2 = 0.5)$	SE
Model I						
β_0	38.62	0.33	38.61	0.28	38.61	0.28
β_1	0.14	0.42	0.16	0.36	0.16	0.36
Model II						
β_0	44.35	1.76	45.39	1.87	51.77	2.54
β_1	0.99	1.91	0.86	2.03	-0.43	2.77
β_2	-2.09	0.25	-2.25	0.27	-3.21	0.38
β_3	-0.11	0.31	-0.09	0.32	0.13	0.45
β_4	0.15	0.01	0.16	0.0	0.14	0.021
Model III						
β_0	45.38	1.54	46.34	1.58	51.85	1.85
β_1	-0.61	1.48	-0.62	1.48	-0.65	1.48
β_2	0.14	0.02	0.14	0.02	0.12	0.03
β_3	0.02	0.03	0.02	0.03	0.02	0.03
β_4	-2.17	0.15	-2.31	0.16	-3.12	0.21

Model I: $PTF = \beta_0 + \beta_1 G + \in,$
Model II: $PTF = \beta_0 + \beta_1 G + \beta_2 PA + \beta_3 PA * G + \beta_4 Age + \in,$
Model III: $PTF = \beta_0 + \beta_1 G + \beta_2 Age + \beta_3 Age * G + \beta_4 PA + \in$

Table 2: Estimates of model coefficients and the standard errors of naive and BC approach for CODING study–Females.

As the tables show, when the impact of ME is very small $(\sigma_u^2 = 0.1)$, Naive and BC approach estimates for the three models are very similar. However, it is not the case when the impact of ME is relatively large $(\sigma_u^2 = 0.5)$. In Model II, Naive estimates of the coefficients for variables *G, PA* and $PA * G$ are affected by the large ME error. Although there was no correction for misclassification of G in BC approach, there is a significant difference between the two estimators of β_1. The reason is the interaction between *G* and the error-prone variable PA. In Model III, however, as it was expected, Naive estimate of the coefficient of the only variable that is highly affected by the ME, is PA. As it was stated in the introduction, the impact of ignoring the ME error, generally, varies from bias in the naive estimators, to false-positive (negative) results in detecting associations. As there is no estimate available for σ_u^2 in this data, it is not possible to find out about the impact. However, some interpretations can be made. The large sample Wald test for all the parameters in Model II in Table 2 indicates that both Naive and BC approaches provide similar significant results, different signs of the estimates for β_1 and β_3 for large variability in ME, however, provides different interpretations of these values. Naive estimates of these parameters imply that for low risk genotype, every additional score in PA makes 4.7% reduction in PTF. For males of high risk genotype, the same amount of increase in PA, obtains only 2.8% reduction in PTF. BC approach, from another hand, starts with higher average PTF for males. It also implies that for males of low risk allele, every additional score in PA makes 6.2% improvement in PTF, when for high risk genotype this amount is 6.8%.

Conclusion

It is now well known that studies of gene-environment interactions can improve the accuracy and precision of the assessment of both genetic and environmental influences. The existing GEI studies on obesity related traits examine both genetics and environmental factors under the assumption that they are measured accurately. However, in reality, both genetics and environmental factors are likely measured with errors. The impact of ignoring errors in variables varies from bias and large variability in estimators to low power or even false negative (positive) results in detecting genetic associations. In order to obtain more accurate results, the bias caused by the errors needs to be corrected.

In this paper, we studied gene-environment interaction and candidate gene association models, where there are misclassification and measurement errors on covariates. In particular, we proposed bias-corrected methods to account for these errors. The proposed methods are easy to apply, and unlike some other bias-corrected methodologies [11], do not require distributional assumptions on ME, and/or error-prone covariates. Our results based on simulation studies show that the proposed methodologies perform quite satisfactory. We also analyzed the CODING data showing that when ME is relatively large, the bias caused by it can dramatically affect the estimation in parameters, and therefore, interpretation of the corresponding values.

There are methodologies suggested by other authors to deal with ME in linear and nonlinear models. Some, studied regression calibration and simulation extrapolation [12-14]. These two methods are only "approximately" consistent, which means that even for large sample size, they still require small ME to perform well. Likelihood-based methods have also been investigated (for example [9] and [12]). Generally, likelihood approaches suffer from restrictive distributional assumptions on ME, covariates with ME and the model error term. Since error-prone covariates and ME are unobservable, likelihood-based approaches might not be realistic. The proposed approaches in this paper do not require parametric assumptions for the distributions of the unobserved covariates and of the measurement errors, which are difficult to check in practice. They also perform well, no matter how large the ME is. Moreover, the same methodologies may be applied to any interaction models between categorical and continuous variables. However, in those models, both sensitivity and specificity are required to be estimated.

ME models, in general, require extra information such as replicate data, internal or external validation data, or instrumental variables, in order to be identifiable. For example, Abarin and Wang [15] proposed a semi-parametric method for estimating parameters of generalized linear regression models with the classical ME model using instrumental variables. In the case that no extra information is available, sensitivity analysis is performed.

The methodology proposed in this paper can be generalized to longitudinal models. Fan et al. [16] proposed a bias-corrected quasi-likelihood approach for longitudinal models, where continuous covariates are subject to error. Generalization of the methodology to longitudinal models, with both misclassified and ME, and the interaction between them, yet to be studied. More studies are also required on the proposed methodology in this paper, to the GEI models where there are more than two categories in the classified variable.

Overall, the results of this paper contribute to enhance the discovery of the genetics and environmental factors in GEI studies. We developed modern yet flexible measurement error techniques that will improve the identification of genetic variants, environmental factors, and their interactions associated with any complex trait.

Acknowledgements

Research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Research and Development Corporation Newfoundland and Labrador (RDC). The author is grateful to Dr. Guang Sun, Faculty of Medicine at Memorial University for providing the data for the analysis. The author is also grateful to the reviewers for their very helpful comments that improved the paper.

References

1. Kaakinen M, Läärä E, Pouta A, Hartikainen AL, Laitinen J, et al. (2010) Life-

course analysis of a fat mass and obesity-associated (FTO) gene variant and body mass index in the Northern Finland Birth Cohort 1966 using structural equation modelling. Am J Epidemiol 172: 653-665.

2. Li S, Zhao JH, Luan J, Ekelund U, Luben RN, et al. (2010) Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. PLoS Med 7: e1000332.

3. Qi L, Cho YA (2008) Gene-environment interaction and obesity. Nutr rev 66: 684-694.

4. Lee SS (2011) Deviant peer affiliation and antisocial behaviour: Interaction with Monoamine Oxidase A (MAOA) genotype. J Abnorm Child Psychol 39: 321332.

5. Fuller AW (2006) Measurement error models. (1st Edn), Wiley-Interscience, New York, USA.

6. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: A modern perspective. (2nd Edn), Chapman & Hall, London, UK.

7. Armstrong BG (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med 55: 651-656.

8. Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ (2003) The detection of gene environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? Int J Epidemiol 32: 51-57.

9. Buonaccorsi JP (2010) Measurement error: models, methods, and applications. (1st Edn), Chapman and Hall/CRC, London, UK.

10. Baecke JA, Burema J, Frijters JE (1982) A short questionnaire for the measurement of habitual physical activity in epidemiological studies. Am J Clin Nutr 36: 936-942.

11. Tosteson TD, Buonaccorsi JP, Demidenko E (1998) Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. Stat Med 17: 1959-1971.

12. Buonaccorsi J, Demidenko E, Tosteson T (2000) Estimation in longitudinal random effects models with measurement error. Stat Sin 10: 885-903.

13. Wang N, Lin X, Gutierrez RG, Carroll RJ (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. J Am Stat Assoc 93: 249-261.

14. Wang N, Lin X, Guttierrez RG (1999) A bias correction regression calibration approach in generalized linear mixed measurement error models. Commun Stat Theory Methods 28: 217-232.

15. Abarin T, Wang L (2012) Instrumental variable approach to covariate measurement error in generalized linear models. Ann Inst Stat Math 64: 475-493.

16. Fan Z, Sutradhar BC, Rao RP (2012) Bias corrected generalized method of moments and generalized quasi-likelihood inferences in linear models for panel data with measurement error. Sankhya B 74: 126-148.