RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Exposure Enriched Case-Control (EECC) Design for the Assessment of Gene–Environment Interaction

Md Hamidul Huque,[1]* Raymond J. Carroll,[1,2] Nancy Diao,[3] David C. Christiani,[3] and Louise M. Ryan[1,4]

[1]*School of Mathematical and Physical Sciences, University of Technology Sydney, New South Wales, Australia;* [2]*Department of Statistics, Texas A&M University, College Station, Texas, United States of American;* [3]*Department of Environmental Health, Harvard School of Public Health, Boston, Massachusetts, United States of American;* [4]*Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of American*

Received 28 December 2015; Revised 7 May 2016; accepted revised manuscript 8 May 2016.
Published online 17 June 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21986

ABSTRACT: Genetic susceptibility and environmental exposure both play an important role in the aetiology of many diseases. Case-control studies are often the first choice to explore the joint influence of genetic and environmental factors on the risk of developing a rare disease. In practice, however, such studies may have limited power, especially when susceptibility genes are rare and exposure distributions are highly skewed. We propose a variant of the classical case-control study, the exposure enriched case-control (EECC) design, where not only cases, but also high (or low) exposed individuals are oversampled, depending on the skewness of the exposure distribution. Of course, a traditional logistic regression model is no longer valid and results in biased parameter estimation. We show that addition of a simple covariate to the regression model removes this bias and yields reliable estimates of main and interaction effects of interest. We also discuss optimal design, showing that judicious oversampling of high/low exposed individuals can boost study power considerably. We illustrate our results using data from a study involving arsenic exposure and detoxification genes in Bangladesh.
Genet Epidemiol 40:570–578, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: Arsenic exposure; case-control; gene–environment; logistic regression; power

Background

Many common diseases are now believed to be the result of interdependence between genetic and environmental factors [Chatterjee and Carroll, 2005; Liu et al., 2012; Mukherjee et al., 2010]. Gene–environment interaction refers to the setting where the effects of an environmental exposure are enhanced in a particular genetic subgroup. Consequently, identification of gene–environment (GE) interactions plays an important role in understanding the aetiology of underlying diseases and hence, developing disease prevention and intervention strategies. However, the classic case-control design can have limited power for studying gene–environment interaction, especially in the case of rare genetic variants and also when exposure distributions are skewed [Foppa and Spiegelman, 1997; García-Closas and Lubin, 1999; Luan et al., 2001]. To address this, various complex sampling strategies have been proposed (see [Thomas, 2010] for a recent review). In one of the first such approaches, White (1982) proposed a two stage design where exposure (or an appropriate surrogate) is first measured in a large number of case and control subjects (Stage I). At Stage II, detailed covariate information is obtained for a subset from each strata defined by case-control and exposure status.

Breslow and Cain (1988) formalized and generalized White's approach to a general two-stage design with analysis proceeding via logistic regression applied to stage II data, but including an offset terms that reflects the stage I sampling probabilities. Weinberg and Wacholder (1990) suggest a slightly simpler approach to the analysis of two stage designs, based on a so-called pseudo likelihood approach that conditions on being sampled in the second stage. Their method also requires inclusion of an offset reflecting sampling probabilities into the logistic regression. While these approaches all provide consistent estimate of main and interaction effects, they require knowledge of the first-stage selection probabilities. In this paper, we propose an alternative approach that does not require knowledge of these probabilities.

Our work is motivated by a study designed to explore the relationship between drinking water arsenic levels, genetic polymorphisms, and skin lesions in Pabna, Bangladesh [Breton et al., 2007]. Because the distribution of arsenic exposure is generally high and right skewed in Bangladesh [Ravenscroft et al., 2005], study investigators had oversampled low exposed individuals (< 50 micrograms per liter) among controls. Consequently, traditional logistic regression analysis was no longer valid, since the sampling mechanism had violated the key assumption for a case-control study, namely that sampling should be independent of exposure status.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Md. Hamidul Huque, School of Mathematical and Physical Sciences, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2207, Australia. E-mail: MdHamidul.Huque@student.uts.edu.au

Our approach is designed for settings where interest lies in characterizing a dose-response relationship and associated interactions based on a continuous exposure. Our exposure enriched case-control (EECC) design oversamples subjects based on case-control status, as well as a categorical assessment of exposure (e.g., high vs. low). We show that as expected, selection of individuals based on high (or low) exposure results in biased estimation of the regression coefficients when standard logistic regression is used. However, we further show that valid statistical inference can be achieved simply by the addition of a single covariate that reflects this exposure-related category. We illustrate via computer simulations that judicious oversampling of individuals based on exposure can significantly boost study power. We also investigate the relative importance of each of the parameters that determine power for detecting interaction effects.

Methods

Suppose the probability of disease occurrence in the general population satisfies a logistic regression model

$$logit\ [\Pr(D = 1|E, G)] = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG, \tag{1}$$

where D = 1 denotes the diseased and D = 0 the nondiseased state, E denotes the level of a continuous environmental exposure, G is a binary indicator of genetic-susceptibility, EG is the gene–environment interaction and where $\beta_0, \beta_E, \beta_G, \beta_{GE}$ are the associated regression coefficients. Genetic susceptibility is defined as the presence of one or more gene mutations thought to be associated with the disease of interest. In practice, the susceptible group will generally correspond to those with the less common variant of the allele of interest [WHO, 2016].

It is well known that while ordinary logistic regression analysis of case-control study data results in incorrect estimation of the intercept (β_0), all other regression coefficients are estimated correctly. This is due to the fact that instead of selecting a random sample from the source population, a biased sample based on case-control status was recruited. There are many approaches to understanding why ordinary logistic regression works, despite the fact that sampling is biased in the case-control setting [Prentice and Pyke, 1979; Weinberg and Wacholder, 1990]. We find it particularly useful to consider a derivation based on Bayes' rule [Hosmer and Lemeshow, 2004]. We use the same principle to show that it is possible to boost study power to detect an interaction effect by oversampling not only cases, but also high (or low) exposed individuals. Similar probabilistic logic was also used by Weinberg and Wacholder [1990].

Define Δ as the sampling indicator with Δ= 1 if the individual is selected into the study sample and 0 otherwise. Also denote the probability of selecting an individual into the sample who has disease status D, exposure level E and genetic characteristic G by $\rho(D, E, G)$ that is, $\rho\ (D, E, G) = \Pr(\Delta = 1|D, E, G)$. Then some simple algebra and an application of Bayes rule leads to the probability of being diseased conditional on exposure, gene and being included in the study sample as (see technical eAppendix A for details).

$$\begin{aligned} &logit\ [\Pr(D = 1|E, G, \Delta = 1)] \\ &= log\left\{\frac{\rho(D = 1, E, G)}{\rho(D = 0, E, G)}\right\} \\ &+\beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG. \end{aligned} \tag{2}$$

Note that the additional term in model (2), compared with model (1), is the log odds of selection for cases vs. controls, conditional on environmental exposure status and genetic susceptibility. This term is associated with the mechanism of recruitment of the study sample, is within the control of investigators and is to be fixed as part of the design. Depending on the recruitment of individuals in the sample, Equation (2) leads to a variety of familiar designs, including:

1. If the sampling probability $\rho(D, E, G)$ is constant, i.e., a simple random sample of subjects is chosen from the population then model (2) will estimate the true population intercept, β_0. This design is popularly known as the prospective cohort study [Prentice and Pyke, 1979].
2. If the sampling probability $\rho(D, E, G)$ depends on the disease status, D but is independent of genetic status (G) or environmental exposure (E), i.e., $\rho\ (D, E, G) = \rho(D)$, then the above model (2) represents an ordinary case-control design with the intercept of the model corresponding to $\beta_0^* = log(\rho(D = 1)/\rho(D = 0)) + \beta_0$. That is, the estimated intercept of the model (2) will be incorrect without the knowledge of the disease prevalence. This result explains the well-known fact that standard logistic regression applied to case-control data yields valid esimates of all regression coefficients except the intercept.
3. If the sampling probability $\rho(D, E, G)$ depends on disease and level of exposure, then its effect on Equation (2) depends on the nature of the relationship. It is Case (3) that we examine in more detail in this paper.

Consider a situation where the selection of individuals depends on a certain cut-off value, *k*, of the observed exposure, E, that characterizes the high or low exposure. Let p_{11} and p_{10} denote the probability of selecting a case in the sample with high (i.e., a subject with D = 1 and E > *k*) and low (i.e., a subject with D = 1 and E < *k*) exposure, respectively. Similarly, let p_{01} and p_{00} denote the probability selecting a control subject in the sample with high (i.e., D = 0 and E > *k*) and low (i.e., D = 0 and E < *k*) exposure, respectively. Then Equation (2) can be reexpressed as

$$logit\,[\Pr(D = 1|E, G, \Delta = 1)] = \begin{cases} log\left(\dfrac{p_{11}}{p_{01}}\right) + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG, & if\ \ E > k \\ log\left(\dfrac{p_{10}}{p_{00}}\right) + \beta_0 + \beta_E E + \beta_G G + \beta_{GE} EG, & if\ \ E < k. \end{cases}$$

This means that the intercept in the model varies according to whether or not E > k. Thus we can succinctly write:

$$logit\ [\Pr(D = 1|E, G, \Delta = 1)]$$
$$= \beta_0^{**} + \lambda I\ [E \geq k] + \beta_E E + \beta_G G + \beta_{GE} EG, \quad (3)$$

where $I\ [E \geq k]$ is an indicator representing whether the environmental exposure E is above the specified level k. The parameter β_0^{**}represents the intercept in the low exposed group and can be expressed as a function of true intercept β_0 and the log odds of selection for cases vs. controls in the low exposed group, i.e., $\beta_0^{**} = log(\frac{p_{11}}{p_{01}}) + \beta_0$. Similarly, λ represents difference between the log odds of the selection of cases vs. controls in the high vs. the low exposed group, i.e., $log(\frac{p_{11}}{p_{01}}) - log(\frac{p_{10}}{p_{00}})$.

Equation (3) suggests a simple approach of adding an indicator variable for high vs. low exposure into the logistic regression on case-control status. Hence, prospective analysis [Prentice and Pyke, 1979] via logistic regression with the addition of this covariate will yield consistent maximum likelihood estimation and inference of regression coefficients associated with exposure, gene, and interaction. Additional covariates can also be included if desired. Throughout this paper, we assume that the sampling depends only on the case-control status and environmental exposure, but is independent of genetic susceptibility.

Simulation Study

In this section, we discuss a simulation study designed to evaluate the finite sample properties of the estimated parameters under the EECC design. We also explore power properties of the proposed methods under various alternatives.

Data Generation

We generated a hypothetical population of one million subjects. A genetic susceptibility covariate, G, was generated as a Bernoulli random variable with prevalence 0.2. The environmental exposure, E, was generated as an exponentially distributed random variable with rate 1. The outcome data were generated using model (1) where parameters β_0, β_{E}, β_{G}, and β_{GE} were set to –4.60, 1.15, 0.8, and 0.406, respectively. The intercept parameter ensures the rare disease assumption with 1% disease prevalence in the control population with nonsusceptible gene. We defined high (or low) exposed individuals if its exposure level was greater (or smaller) than an exposure level of 2, which corresponds to having 13.5% of the exposure data in the upper tail area. We then selected a total sample of 1100 observations from the above population equally stratified by exposure level and case status. An equal number of high and low exposed subjects in the sample thus result in oversampling from high exposed group.

Parameter Estimation

We estimated relevant parameters of our proposed method applying logistic regression with an indicator variable (3) to the data generated according to the scheme described above. We also compare these estimates with a naive analysis that excludes the indicator variable. The sampling and estimation procedure was repeated 1000 times.

Power Calculation

We use a simulation based technique to calculate power for testing $H_0: \beta_{GE} = 0$ vs. $H_A: \beta_{GE} \neq 0$ via a standard Wald test [Cox and Hinkley, 1974]. To estimate the power of the test, we simulated data under the alternative hypothesis, H_A and EECC design, fitted model (3). Again, there were 1,000 simulated datasets. The estimated power is the proportion of the 1000 replicates whose test statistics exceeds the relevant critical value of ±1.96 (at 5% level of significance). Though we use 5% level of significance, a smaller level of significance can also be incorporated in testing the hypothesis. All calculations were performed using **R** [RCoreTeam 2014].

Simulation Results

We first show that ignoring the sampling scheme and performing standard logistic regression results in biased estimation. We then show that the bias can be removed through addition of an indicator variable, indicating high exposure, in the logistic regression model (3). We later calculate power of our proposed method by varying different parameters that govern the power to detect gene–environment interaction.

Parameters Estimation

Figure 1 compares the distribution of the estimated regression coefficients using logistic regression ignoring the over sampling of high exposed subjects (Fig. 1A) and accounting for high exposed subject in the sample (Fig. 1B). In this case, data were generated according to the EECC design and using the parameters values describe in the data generation section above. The dotted lines indicate the true value of the corresponding parameters. As the boxplots indicate, performing a standard logistic regression results in incorrect estimates of all the parameters of the logistic regression model (Fig. 1A). However, adding an extra covariate indicating high exposure yields reliable estimates of the true parameters (Fig. 1B).

We also conducted additional simulations to exposure the nature of the bias in estimated regression coefficients when there is no interaction ($\gamma = 0$) or a negative interaction ($\gamma = -0.406$) in the true model. The results are given in eFigure 1. The results are quite similar to the results with positive interaction parameters, i.e., traditional analysis ignoring sampling provides biased estimates, however, analysis by adding an indicator variable in the model provides reliable estimates of the true parameters.

We further conduct a simulation study similar to the data collection in our motivating example where cases were randomly selected without stratification and low exposed controls are oversampled. Specifically, we set a lower cut off value (of $k = 0.4$, which for the exponential exposure

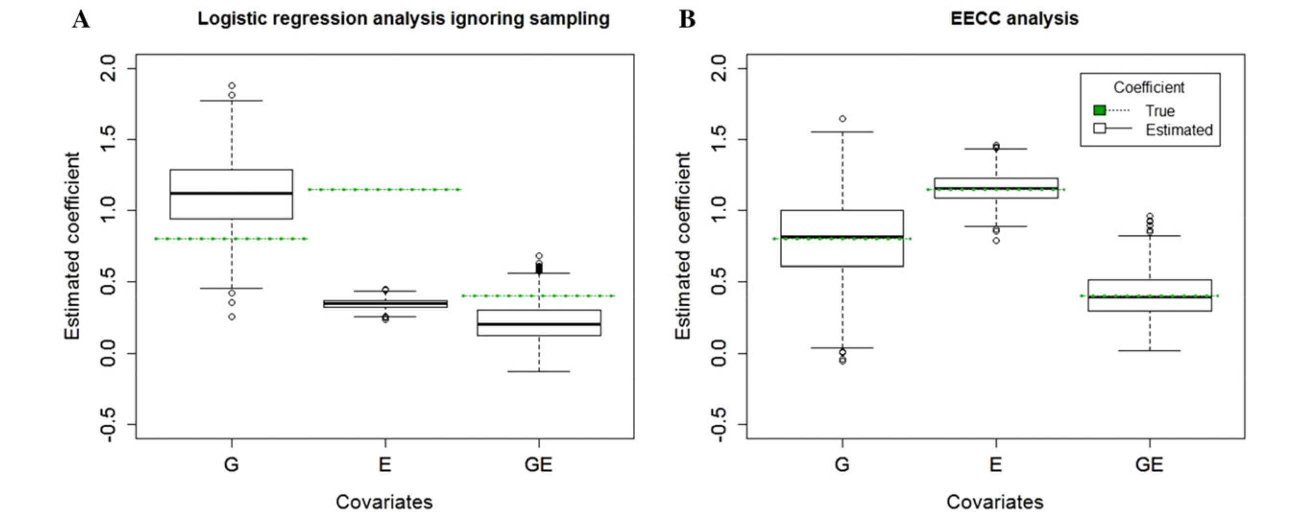

Figure 1. Comparison of estimated coefficients obtained using usual logistic regression ignoring sampling and proposed method.

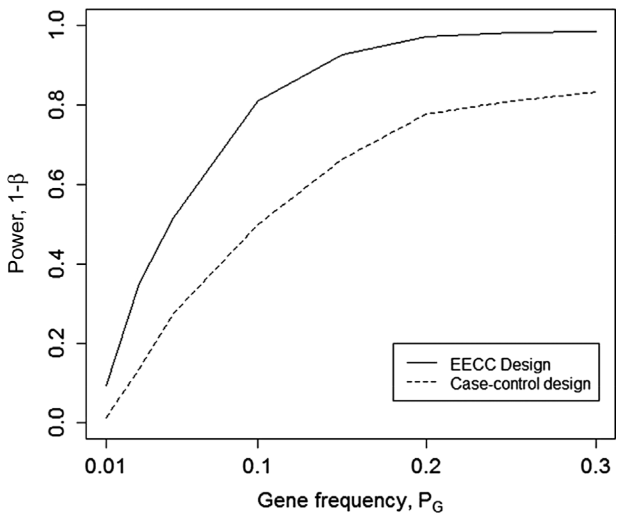

Figure 2. Comparison of estimated power to detect gene–interaction effect for a sample size of 2,000 obtained using traditional case-control design and EECC design.

distribution corresponds to 33% being in the low exposed group. We then sampled equal number of controls from high/low exposed group based on above cut-off. All other parameters involved in this simulation were similar to values presented in Figure 1. The results are given in eFigure 2. The EECC design in this case also provides reliable estimates of the true regression coefficients.

Power Estimation

In this section, we will compare power to detect gene–environment interaction effect employing EECC design and traditional case-control design (sampling from case and control population independent of exposure status). Given a particular value of the gene–environment interaction parameters under the alternative hypothesis, the power is a function of the following parameters: the magnitude of the type I error (α), the sample size (n), the gene frequency (P_G), the exposure distribution (E), the control to case ratio(r_c), the ratio of high and low exposed sample(r_H), and the cut-off, k above (or below) which the exposure is considered to be high (or low). Therefore, we estimate power by varying one of the aforementioned parameters, the remaining parameters were held fixed. For all the power comparisons, unless stated otherwise the exposure distribution is considered as exponential (rate = 1) and distribution of gene prevalence is considered as Binomial (P_G = 0.2).

Relation Between Power and Gene Frequency

Figure 2 illustrates the power comparison to detect the interaction parameter β_{GE} using traditional case-control design and EECC design for different values of the prevalence of genetic susceptibility, P_G and a total sample size of 2000. As expected, power to detect interaction parameter decreases with low disease prevalence. However, the EECC design yields better power compare to traditional design for all cases with various probability of gene susceptibilities.

Relation Between Power and Case-Control Ratio

In Figure 3, the power is shown as a function of control-case ratio (r_C) for a given number of cases using EECC design, see Figure 3A, and using traditional case-control design; see Figure 3B. Power increases as the number of controls increases for a fixed number of cases in both the design. Similar to the classical case-control studies [Taylor 1986], most of the gain is evident if the control to case ratio is at most 4 in EECC design. However, the EECC design outperforms the

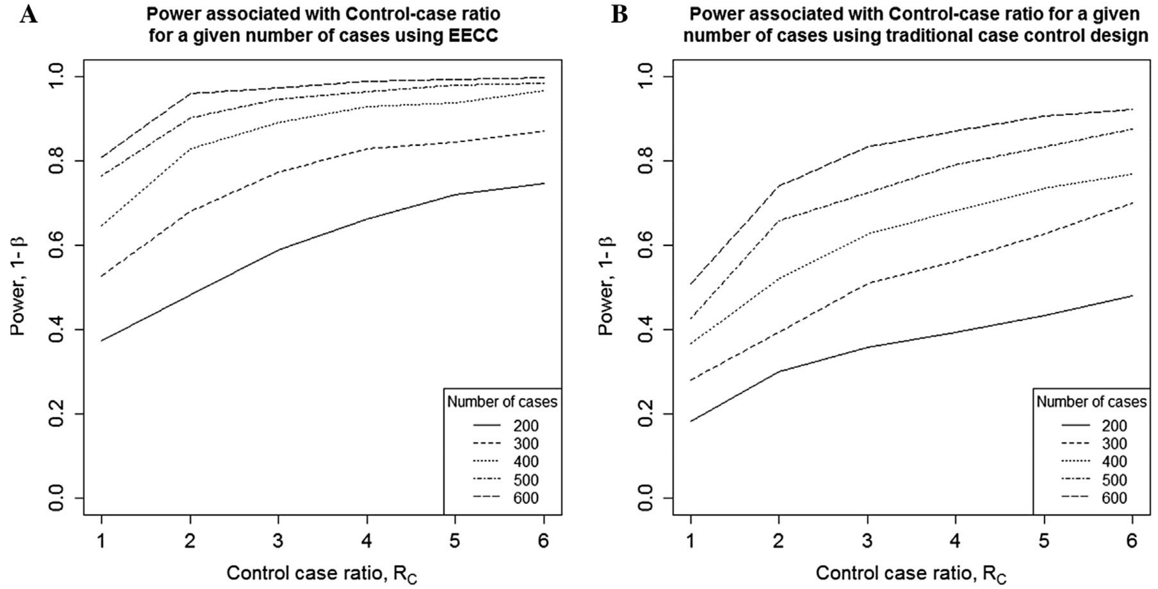

Figure 3. Power as a function of case-control ratio and sample sizes: (A) EECC design (B) traditional case-control design.

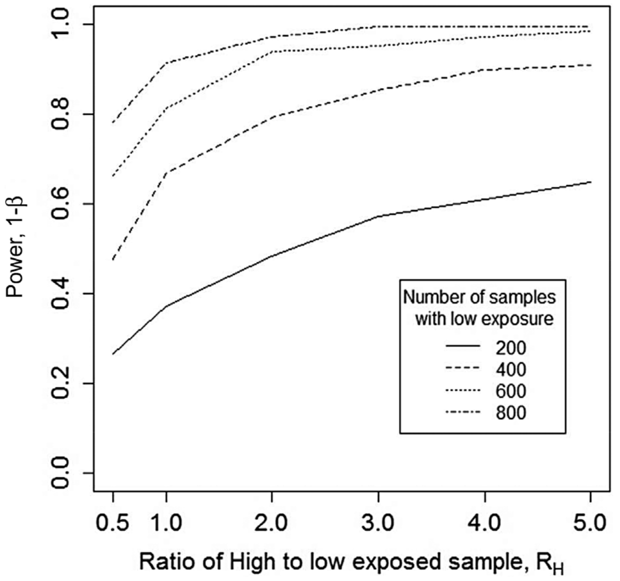

Figure 4. Power as a function of ratio of high exposed sample compared with low exposed sample with varying number of low exposed samples.

traditional case-control design in obtaining power to detect gene–environment interaction.

Relation Between Power and Ratio of High and Low Exposed Sample

In Figure 4, the power is shown as a function of high to low exposure ratio (r_H) in the sample with equal number of case and control. Sampling of more high exposed subjects compared to low exposed subjects resulted in increased power for exponential (rate = 1) distribution. Most of the gain in terms of power is achieved if the ratio of high to low exposed subjects is between 1 and 3.

Relation Between Power and Asymmetry of the Exposure Distribution

In Table 1, we evaluate the estimated regression coefficients, their standard errors and power for detecting the gene–environment interaction effect as a function of the asymmetry of the exposure distribution and level of exposure cut off, k. Various values of cut off were examined to ensure varying proportions (5–95%) of exposure information lies above (or below) the cut off. Specifically, we simulated exposure distribution following Beta (6, 2), Beta (6, 6), and Beta (2, 6), where beta (α, β) represents Beta distribution with shape parameters α and β. Under the above specification the exposure distribution is either negatively skewed, symmetric, or positively skewed. The true parameter values used in this particular simulation is given in the first row of the Table 1. As expected, the power to detect a significant interaction effect increases considerably with oversampling from the skewed tail area of the exposure distribution. In the case of a symmetric exposure distribution, oversampling from lower tail areas boosts in power. The estimated regression coefficients remain consistent to the true values. Furthermore, if the cut off for oversampling is selected appropriately, there is a gain in efficiency for estimating gene and gene–environment interaction effect. However, addition of an indicator variable related to exposure status decrease efficiency for the continuous exposure effect.

Table 1. Comparison of estimated regression coefficients and power for asymmetry of the exposure distribution and varying cut offs with a sample size of 1,600

			Traditional case-control design				Modified case-control design			
Exposure distribution	% of exposure in the left tail	Cut off	$\hat{\beta}_E$ $se(\hat{\beta}_E)$	$\hat{\beta}_G$ $se(\hat{\beta}_G)$	$\hat{\beta}_{EG}$ $se(\hat{\beta}_{EG})$	Power	$\hat{\beta}_E$ $se(\hat{\beta}_E)$	$\hat{\beta}_G$ $se(\hat{\beta}_G)$	$\hat{\beta}_{EG}$ $se(\hat{\beta}_{EG})$	Power
True coefficients			**1.15**	**0.80**	**1.5**		**1.15**	**0.80**	**1.5**	
Beta (6,2) (Left skewed)	5%	0.48	1.167 (0.533)	0.830 (0.711)	1.480 (0.920)	0.372	1.132 (0.658)	0.793 (0.376)	1.153 (0.603)	0.729
	10%	0.55					1.181 (0.687)	0.823 (0.465)	1.487 (0.708)	0.592
	20%	0.63					1.175 (0.693)	0.803 (0.506)	1.511 (0.736)	0.503
	30%	0.69					1.127 (0.783)	0.792 (0.593)	1.518 (0.829)	0.439
	40%	0.73					1.164 (0.753)	0.825 (0.660)	1.476 (0.885)	0.406
	50%	0.77					1.168 (0.782)	0.826 (0.693)	1.481 (0.899)	0.391
	60%	0.81					1.161 (0.793)	0.799 (0.763)	1.519 (0.974)	0.393
	70%	0.84					1.187 (0.768)	0.814 (0.775)	1.492 (0.966)	0.383
	80%	0.88					1.133 (0.738)	0.784 (0.751)	1.527 (0.900)	0.380
	90%	0.92					1.145 (0.719)	0.801 (0.730)	1.509 (0.857)	0.381
	95%	0.95					1.128 (0.693)	0.774 (0.785)	1.543 (0.904)	0.410
Beta (6,6) (Symmetric)	5%	0.27	1.178 (0.502)	0.823 (0.469)	1.474 (0.886)	0.363	1.182 (0.650)	0.799 (0.279)	1.516 (0.678)	0.595
	10%	0.32					1.170 (0.679)	0.792 (0.337)	1.540 (0.762)	0.534
	20%	0.38					1.167 (0.722)	0.796 (0.377)	1.527 (0.835)	0.450
	30%	0.42					1.134 (0.705)	0.767 (0.405)	1.562 (0.845)	0.442
	40%	0.46					1.183 (0.732)	0.814 (0.450)	1.488 (0.904)	0.375
	50%	0.50					1.120 (0.763)	0.780 (0.471)	1.555 (0.904)	0.400
	60%	0.54					1.131 (0.739)	0.793 (0.470)	1.531 (0.868)	0.393
	70%	0.58					1.178 (0.769)	0.801 (0.531)	1.516 (0.918)	0.399
	80%	0.62					1.160 (0.726)	0.801 (0.500)	1.510 (0.851)	0.430
	90%	0.68					1.143 (0.716)	0.806 (0.495)	1.502 (0.792)	0.460
	95%	0.73					1.162 (0.652)	0.833 (0.491)	1.459 (0.751)	0.499
Beta (2,6) Right skewed	5%		1.156 (0.448)	0.802 (0.255)	1.531 (0.851)	0.433	1.162 (0.605)	0.800 (0.173)	1.532 (0.780)	0.498
	10%	0.08					1.139 (0.640)	0.786 (0.193)	1.601 (0.838)	0.471
	20%	0.12					1.149 (0.660)	0.800 (0.191)	1.512 (0.832)	0.428
	30%	0.16					1.162 (0.685)	0.797 (0.235)	1.525 (0.908)	0.446
	40%	0.19					1.181 (0.670)	0.799 (0.228)	1.555 (0.836)	0.420
	50%	0.23					1.147 (0.675)	0.799 (0.243)	1.523 (0.835)	0.440
	60%	0.27					1.152 (0.699)	0.786 (0.265)	1.562 (0.825)	0.467
	70%	0.31					1.205 (0.662)	0.814 (0.277)	1.484 (0.799)	0.468
	80%	0.37					1.130 (0.692)	0.806 (0.284)	1.508 (0.744)	0.518
	90%	0.45					1.147 (0.648)	0.800 (0.301)	1.519 (0.715)	0.619
	95%	0.52					1.146 (0.593)	0.805 (0.290)	1.512 (0.626)	0.701

Table 2. Comparison of estimated coefficients their standard errors, power of traditional case-control design vs. modified case-control design for various sample size and exposure distribution

			Traditional case-control design				Modified case-control design			
Sample size, n	exposure distribution	Cut off	$\hat{\beta}_E$ $se(\hat{\beta}_E)$	$\hat{\beta}_G$ $se(\hat{\beta}_G)$	$\hat{\beta}_{EG}$ $se(\hat{\beta}_{EG})$	Power	$\hat{\beta}_E$ $se(\hat{\beta}_E)$	$\hat{\beta}_G$ $se(\hat{\beta}_G)$	$\hat{\beta}_{EG}$ $se(\hat{\beta}_{EG})$	Power
True regression coefficients			1.15	0.80	0.406		1.15	0.80	0.406	
1600	Exp (1)	2.30	1.149	0.780	0.418	0.654	1.153	0.797	0.408	0.938
			(0.072)	(0.253)	(0.181)		(0.090)	(0.241)	(0.119)	
	Weibull (1.5, 1)	1.74	1.152	0.786	0.430	0.499	0.149	0.791	0.425	0.761
			(0.099)	(0.258)	(0.220)		(0.130)	(0.279)	(0.170)	
	Gamma (3,2)	2.66	1.155	0.784	0.424	0.574	1.156	0.810	0.408	0.838
			(0.079)	(0.335)	(0.198)		(0.100)	(0.332)	(0.143)	
1200	Exp (1)	2.30	1.156	0.781	0.427	0.530	1.158	0.794	0.412	0.861
			(0.082)	(0.291)	(0.215)		(0.097)	(0.289)	(0.141)	
	Weibull (1.5,1)	1.74	1.155	0.769	0.437	0.405	1.156	0.792	0.415	0.620
			(0.120)	(0.300)	(0.254)		(0.148)	(0.311)	(0.188)	
	Gamma (3,2)	2.66	1.156	0.789	0.424	0.459	1.161	0.811	0.409	0.725
			(0.097)	(0.397)	(0.233)		(0.121)	(0.371)	(0.163)	

Exponential with rate 1, Weibull (shape 1.5, scale = 1), and Gamma (shape = 3, rate = 2).
The various cut off ensures 10% of the total exposure data lies above these cut offs.

Relation Between Power and Exposure Distributions

To examine the performance of our proposed method with other exposure distribution, we compared the performance of EECC design with traditional case-control design for various right tailed exposure distributions, e.g., exponential (rate = 1), Weibull (shape = 2.5, scale = 1), and gamma (shape = 3, rate = 2). We select suitable cut off values so that a 10% of the total exposure information lies above these cut off. The results are given in the Table 2. For all the sample sizes and exposure distribution compared, the EECC design resulted in higher power for detecting the interaction effect than the traditional case-control study design.

Relation Between Power/Probability of Type I Error With the Signs of Interaction Parameter

We also estimate the power and probability of type I error in simulation studies corresponding to the true interaction parameters –0.406 and 0, respectively with exponential exposure distribution. The power to detect negative interaction parameter remains similar to that of positive interaction parameters with the same cut off for oversampling. This indicates that power does not depend on the sign of the interaction parameters rather the skewness of the true exposure distribution. Moreover, type I error probability corresponding to testing interaction parameter 0, for the EECC design remained closed to 0.05 (results not shown in table).

Application to the Arsenic Exposure Data From Bangladesh

We reanalyze data from a case-control study designed to evaluate the joint effect of genetic polymorphisms and drinking water arsenic exposure on skin lesions [Breton et al., 2007]. These data were collected from 23 villages of the Pabna district in Bangladesh, where a range of high and low well water arsenic levels were suspected due to their proximity to the Ganges river. Therefore, to ensure a sufficient range of drinking water arsenic exposure and to prevent overmatching on exposure, the study investigators made sure that 80% of the controls were selected from communities having suspected low exposed arsenic contamination $(< 50\mu_g/l)$. More detailed descriptions of the data collection have been given elsewhere [Breton et al., 2007; McCarty et al., 2006].

Previously, analyzing these data, Breton et al. (2007), reported that the X-ray repair cross complementing group 1 (XRCC1 Arg194Trp) polymorphism has a significant interaction with toenail arsenic concentrations. We evaluate this relationship employing our proposed method. We defined an indicator variable that indicates whether or not the sample is obtained from high exposed communities (≥ 50μ_g/l). We assessed the gene–environment interaction in a crude and adjusted logistic regression model, with the latter accounting for the potential confounders such: age, sex, village, body mass index (BMI), education, ever smoked status, and ever chewed betel nuts status. These models were then compared with modified crude and adjusted logistic regression model (3).

Of the 1,800 participating cases and controls, 1,756 (98%) were genotyped successfully for XRCC1 Arg194Trp genotypes. Complete information on well water arsenic, toenail arsenic, BMI, education, smoking, and betel nut chewing was available for 1,676 (839 controls and 837 cases) of these participants. The resulting data were analyzed using both the traditional logistic regression analysis and EECC. Crude and adjusted odds ratios along with 95% CI and *P*-values for both methods are displayed in Table 3. The results are qualitatively similar in that we see a significant association between arsenic exposure and skin lesions, as well as a significant gene–environment interaction. Specifically, participants with the Trp/Arg genotype have a significantly stronger dose response associated with arsenic exposure. However, the magnitudes of estimated effects are different between the traditional and EECC analyses. Whereas the traditional analysis suggested an adjusted odds ratio of 4.06 (95% CI: 3.14–5.25),

Table 3. Comparison of results obtained by employing traditional and EECC design to the arsenic data

	Traditional design						EECC design					
	Crude analysis			Adjusted analysis			Crude analysis			Adjusted analysis		
Exposure	OR	95% CI	*P*-value	aOR	95% CI	*P*-value	OR	95% CI	*P*-value	aOR	95% CI	*P*-value
log(Toenail)	3.99	(3.17–5.27)	<0.001	4.06	(3.14–5.25)	<0.001	3.10	(2.35–4.09)	<0.001	3.15	(2.37–4.20)	<0.001
XRCC1Arg194Trp												
Arg/Arg	Ref			Ref			Ref			Ref		
Trp/Arg	1.09	(0.76–1.48)	0.61	1.08	(0.79–1.48)	0.62	1.08	(0.79–1.48)	0.63	1.08	(0.79–1.47)	0.65
Trp/Trp	1.21	(0.44–3.33)	0.71	1.29	(0.46–3.58)	0.63	1.22	(0.44–3.35)	0.70	1.29	(0.46–3.61)	0.63
log(Toenail) * *XRCC1Arg194Trp*												
lnTA*Trp/Arg	0.53	(0.32–0.91)	0.02	0.53	(0.31–0.90)	0.02	0.56	(0.33–0.94)	0.03	0.55	(0.33–0.94)	0.03
lnTA*Trp/Trp	0.30	(0.05–1.66)	0.17	0.27	(0.06–2.07)	0.14	0.28	(0.05–1.58)	0.15	0.26	(0.05–1.45)	0.12

our EECC analysis reduced the estimated dose-response coefficient to 3.15 (95% CI: 2.37–4.20). The EECC analysis has much greater efficiency (smaller CI lengths) compare to traditional design. The estimates for genetic effect and gene–environment interaction effect are similar for the traditional and EECC design.

Discussion

In this paper, we have introduced the EECC study that oversamples according to case-control status, as well as a categorization of exposure (e.g., high vs. low). We have shown via simulations that the EECC can significantly boost the power to detect gene–environment interaction, especially in the case of rare genetic variants and skewed exposure distributions. Stenzel et al. (2015) also use the term "Exposure Enriched" and argue that oversampling high exposure can boost the power to detect gene–environment interactions [Stenzel et al., 2015]. However, they analyzed the resulting data via ordinary methods and did not suggest any analysis strategy to remove the biased induced by oversampling highly exposed individuals. Our EECC method removes the bias induced by oversampling high exposed individuals through the addition of a simple indicator covariate that reflects high vs. low exposure. Our approach assumes that case-control status will be modeled as a function of a continuous exposure variable, genetic susceptibility, and their interactions. Our approach differs from other analysis methods for data collected via biased sampling in that it does not require knowledge of the explicit sampling probabilities [Breslow and Cain, 1988; Weinberg and Wacholder, 1990; White, 1982]. Our proposed EECC method has the advantage of simplicity since no specialized software is required.

Although existing two stage case-control designs [Breslow and Cain, 1988] and their matched variant, counter matching [Andrieu et al., 2001], are known to have higher power than traditional case-control design, they can only be used if surrogate information on gene, exposure or both is available. The efficiency obtained from these two designs though similar, counter matching designs are complex and require two specific and sensitive surrogates for the risk factor of interest [Andrieu et al., 2001]. Our EECC design is simpler and use similar underlying probability principle as pseudo-likelihood analysis based on a two stage design, hence will results in similar efficiency for an appropriate oversampling of high exposed individuals. Other designs such as family-based designs (see [Thomas, 2010] for a recent review) are appealing in gene–environment interaction studies. However, they generally have less power to test main effects, relative to case-control studies using unrelated controls [Thomas, 2010]. Moreover, they are very sensitive to the independence assumptions of gene and environment effects [Albert et al., 2001]. The empirical comparison of the above designs with our proposed EECC design is beyond the scope of the current paper. However, interested reader might consider recent review [Thomas, 2010] for a detailed comparison among some of these methods.

While our paper has focussed primarily on introducing the EECC method, we have also presented a reanalysis of data from a case-control study from Bangladesh, where low exposed control subjects had been oversampled [McCarty et al., 2006]. McCarty et al. (2006) had used traditional logistic regression with a sensitivity analysis to explain the effect of this biased sampling. However, the authors reported that they were not able to make a succinct conclusion about the observed exposure-response relationship between arsenic levels in tube-well drinking water and skin lesions, due to oversampling of controls from the low exposed area. Our EECC approach rectifies the analysis with the addition of an extra covariate indicating the oversampling rules in the model.

Although our proposed EECC methodology has a number of appealing features, there are some limitations that could be addressed in future studies. We assumed a linear relationship between the log-odds of disease and exposure. Additional simulation studies suggest that misspecification of this assumption will produce bias results for both traditional case-control studies and EECC designs (results not shown in the Table). It should be straightforward to relax the linearity assumption for the EECC design and the method will work so long as we assume a smooth relationship between exposure and the log-odds of disease. Essentially the EECC design exploits the fact that a discontinuity in the exposure response relationship is induced by oversampling individuals with exposure levels above a cut-off value *k*.

In practice, there may be limitations in terms of the availability of subjects who meet a specified selection criteria. For example, suppose a study design aims to recruit equal numbers of case and control subjects in both high and low exposure categories. If the cut-off is set too high (or low), then there may not be enough high (or low) exposure subjects available. Therefore, one might need to adjust the cut-off values to ensure that the design is feasible.

Our proposed EECC design is currently allows only a binary cut off variable to represent oversampling from tail area. However, in application more than one exposure levels in the tail area might be of interest. Future work need to accommodate such extension. In environmental epidemiology, exposure is often susceptible to measurement error [Huque et al., 2014]. In the case of exposure misclassification, it is well known that the estimates of the regression coefficients will be attenuated [Stefanski and Carroll, 1985] and may distorts the power gain of exposure enriched design. Although various methods have been proposed in the literature to correct the effect exposure measurement error in gene–environment interaction studies [Lobach et al., 2010; Lobach et al., 2011; Spiegelman et al., 2000; Zhang et al., 2008], however, further research is needed to evaluate and incorporate such extension into our proposed EECC methodology.

Despite these potential limitations, our EECC design can be regarded as a simple alternative to traditional two-stage designs. Furthermore the EECC methodology enhances power to detect the joint influence of genetic and environment exposure for a given sample size compare to traditional case-control studies. Therefore, it has a very strong potential to be used in practice. This design also has potential to be used in context of risk analysis where interest lies in quantifying dose response relationships [Piegorsch, 2010].

Supplementary Materials

Supplementary digital content e.g., eAppendix A, eFigure 1 and eFigure 2 is available with this manuscript.

Acknowledgments

Huque and Ryan were supported by the University of Technology Sydney and by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). Carroll was supported by a grant from the National Cancer Institute (U01-CA057030). Christiani and Diao were supported by grants from the National Institute of Environmental Health Sciences (ES-004957, ES-011622 and ES-000002).

References

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene–environment interactions. *Am J Epidemiol* 154(8):687–693.

Andrieu N, Goldstein A, Thomas D, Langholz B. 2001. Counter-matching in studies of gene–environment interaction: efficiency and feasibility. *Am J Epidemiol.* 153(3):265–274.

Breslow N, Cain K. 1988. Logistic regression for two-stage case-control data. *Biometrika* 75(1):11–20.

Breton CV, Zhou W, Kile ML, Houseman EA, Quamruzzaman Q, Rahman M, Mahiuddin G, Christiani DC. 2007. Susceptibility to arsenic-induced skin lesions from polymorphisms in base excision repair genes. *Carcinogenesis* 28(7):1520–1525.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92(2):399–418.

Cox DR, Hinkley DV. 1974. *Theoretical Statistics.* Florida, USA: CRC Press/Chapman & Hall.

Foppa I, Spiegelman D. 1997. Power and sample size calculations for case-control studies of gene–environment interactions with a polytomous exposure variable. *Am J Epidemiol* 146(7):596–604.

García-Closas M, Lubin JH. 1999. Power and sample size calculations in case-control studies of gene–environment interactions: comments on different approaches. *Am J Epidemiol* 149(8):689–692.

Hosmer DW, Lemeshow S. 2004. *Applied logistic regression.* New York: John Wiley & Sons.

Huque MH, Bondell H, Ryan L. 2014. On the impact of covariate measurement error on spatial regression modelling. *Environmetrics* 25(8):560–570.

Liu C, Maity A, Lin X, Wright R, Christiani D. 2012. Design and analysis issues in gene and environment studies. *Environ Health* 11(1):1–15.

Lobach I, Fan R, Carroll RJ. 2010. Genotype-based association mapping of complex diseases: gene–environment interactions with multiple genetic markers and measurement error in environmental exposures. *Genet Epidemiol* 34(8):792–802.

Lobach I, Mallick B, Carroll RJ. 2011. Semiparametric Bayesian analysis of gene–environment interactions with error in measurement of environmental covariates and missing genetic data. *Stat Its Interface* 4(3):305–316.

Luan J, Wong M, Day N, Wareham N. 2001. Sample size determination for studies of gene–environment interaction. *Int J Epidemiol* 30(5):1035–1040.

McCarty KM, Houseman EA, Quamruzzaman Q, Rahman M, Mahiuddin G, Smith T, Ryan L, Christiani DC. 2006. The impact of diet and betel nut use on skin lesions associated with drinking-water arsenic in Pabna, Bangladesh. *Environ Health Persp* 114(3):334–340.

Mukherjee B, Ahn J, Gruber SB, Ghosh M, Chatterjee N. 2010. Case-control studies of gene–environment interaction: Bayesian design and analysis. *Biometrics* 66(3):934–948.

Piegorsch WW. 2010. Translational benchmark risk analysis. *J Risk Res* 13(5):653–667.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66(3):403–411.

Ravenscroft P, Burgess WG, Ahmed KM, Burren M, Perrin J. 2005. Arsenic in groundwater of the Bengal Basin, Bangladesh: distribution, field relations, and hydrogeological setting. *Hydrogeol J* 13(5–6):727–751.

R CoreTeam. 2014. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Spiegelman D, Rosner B, Logan R. 2000. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J Am Stat Assoc* 95(449):51–61.

Stefanski LA, Carroll RJ. 1985. Covariate measurement error in logistic regression. *Annal Stat* 13(4):1335–1351.

Stenzel SL, Ahn J, Boonstra PS, Gruber SB, Mukherjee B. 2015. The impact of exposure-biased sampling designs on detection of gene–environment interactions in case-control studies with potential exposure misclassification. *Eur J Epidemiol* 30(5):413–423.

Taylor JM. 1986. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Stat Med* 5(1):29–36.

Thomas D. 2010. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet* 11(4):259–272.

Weinberg CR, Wacholder S. 1990. The design and analysis of case-control studies with biased sampling. *Biometrics* 46(4):963–975.

White JE. 1982. A two stage design for the study of the relationship between a rare exposure and a rare disease. *AmJ Epidemiol* 115(1):119–128.

WHO. 2016. Genes and noncommunicable diseases. *Genes Hum Dis* http://www.who.int/genomics/public/geneticdiseases/en/.

Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V. 2008. Accounting for error due to misclassification of exposures in case–control studies of gene–environment interaction. *Stat Med* 27(15):2756–2783.