Published in final edited form as:

Epidemiology. 2011 March ; 22(2): 262–272. doi:10.1097/EDE.0b013e318207ffc3.

Robust Detection of Genetic Associations in the Presence of Gene-Environment Interaction and Independence

Eric J. Tchetgen Tchetgen

Departments of Epidemiology and Biostatistics, Harvard University

Abstract

We consider the detection and evaluation of genetic effects in the presence of gene-environment interaction and independence. Whereas ordinary logistic regression cannot exploit the gene-environment independence assumption, we present a novel approach which makes explicit use of this assumption to improve detection and evaluation accuracy of genetic associations. For a binary genetic variant, our method which uses both cases and controls, entails fitting a constrained retrospective logistic regression, in which the genetic variant plays the role of the response variable, and both the disease indicator and the environmental exposure are the independent variables. The regression model constrains the parameters corresponding to the environmental exposure odds ratio association with the genetic variant among the controls to be null, thus explicitly encoding the gene-environment independence assumption. The main appeal of the proposed retrospective regression approach is that (i) it is easy to implement with standard software, (ii) it readily accounts for multiple environmental exposures of a polytomous or of a continuous nature, and also allows for any additional observed covariates, (iii) it is more efficient than ordinary logistic regression (iv) unlike ordinary logistic regression, and the profile likelihood approach of Chatterjee and Carroll (2005), our approach does not require a model for the odds ratio association of the environmental exposure with the disease outcome; as a result it is agnostic and thus completely robust to possible model mis-specification which can happen with an exposure of a polytomous or of a continuous nature

1 Introduction

It is widely recognized that interactions between genetic and environmental risk factors play an important role in the development of many complex diseases. As a result, research initiatives to identify novel interactions between genetic variants and environmental exposures are fast growing, with the hope that such findings will lead to a better understanding of disease etiology and improved treatment and prevention guidelines. Some of the most recent statistical developments to facilitate the detection of gene-environment interaction have exploited a crucial assumption that genetic and environmental factors are distributed independently of each other in the underlying population. In particular, the case-only estimator of an interaction on the risk ratio scale is valid only if this independence assumption is correct. For binary genetic and environmental factors, G and E , the case-only

estimator is the empirical marginal odds ratio (i.e. crude odds ratio) between G and E among cases (with disease status $D = 1$), and thus data on unaffected individuals (with disease status $D = 0$) is not required (Piegorsch et al, 1994). When the disease is rare within all levels of G and E in the population, the case-only estimator is well known to also accurately estimate a $G \times E$ interaction on the odds ratio scale. In addition, efficiency considerations have also contributed to the appeal of the case-only estimator which is nearly efficient when the disease is rare. This is true even when data on unaffected individuals are available as we assume hereafter, by considering the setting of a case-control study with controls sampled from subjects who remain unaffected at the end of the study. In contrast, under gene-environment independence, the standard prospective logistic regression estimator of gene-environment interaction can be less efficient. This is expected since it does not make use of the assumed independence (Prentice and Pyke, 1979; Breslow et al 2000). While the case-only design allows for estimation of G-E interaction, it does not allow for the estimation of the main effects of either G or E on the risk of disease, without which a meaningful interpretation of an interaction may be difficult. Under a rare disease assumption, Modan et al (2001) in a specific application, noted that, under gene-environment independence, the disease odds ratio associated with G among subjects with exposure level $E = e$ can be estimated by a logistic regression analysis that compares the distribution of G among all controls $\Pr\{G|D=0\}$, with the genotype distribution among cases with $E = e$, $\Pr\{G|D=1, E=e\}$. A similar approach gives the disease odds ratio associated with E among subjects with exposure level $G = g$ by simply reversing the roles of gene and environment in the previous description.

In a more general approach, Umbach and Weinberg (1997) obtained maximum likelihood estimates of main effect and interaction parameters of genetic and environmental factors under gene-environment independence, by fitting a suitably constrained log-linear model to the data. Further generalizing these methods, Chatterjee and Carroll (2005) recently proposed a profile maximum likelihood approach which exploits the independence assumption to provide efficient estimates of main effects of E and G and of their interaction. These two methods are particularly appealing because they can be extended to accommodate additional covariates in the model. Additional covariates will often be considered for modeling to control for confounding, as well as to guarantee that gene-environment independence holds within strata of the covariates in the underlying population (Chatterjee and Carroll 2005). The assumption of conditional independence is a weaker and often more appropriate assumption than the assumption of gene-environment marginal independence in the underlying population. Unfortunately, in addition to a model for the disease risk regression $\Pr(D|G, E, L)$, both of these methods also require estimating a distribution of the covariates. On the one hand, the approach of Umbach and Weinberg (1997) requires an estimate of the joint distribution of (G, E, L) , and therefore may not be practical when the set of covariates is of moderate to large size, or contains continuous covariates (Chatterjee and Carroll, 2005). On the other hand, the profile likelihood approach requires a consistent estimate of the conditional density of the genetic factor given L in the underlying population. For a rich model with several covariates, nonparametric estimates of this density will generally perform poorly in finite samples. For this reason, Chatterjee and Carroll (2005)

propose the use of parsimonious parametric models for estimating both $\Pr(D|G, E, L)$ and $\Pr(G=g|L)$.

The extent to which model mis-specification of the main effect of an environmental exposure on the risk of disease can negatively impact inferences on genetic effects based on either standard prospective logistic regression or the profile likelihood approach, has only recently received serious consideration (Cornelis et al. (2010), Tchetgen Tchetgen and Kraft (2010)). In this paper, we briefly review specific conditions given by Tchetgen Tchetgen and Kraft (2010), under which robust inferences to this form of model mis-specification, can be recovered for genetic effects. In addition, we describe an alternative approach to available methods in the form of a simple and more robust retrospective regression framework for the evaluation of genetic effects incorporating gene-environment interaction and independence. In the simple setting where E is binary, G is coded as a binary genetic variant and L is absent, our approach consists of using both cases and controls in a constrained retrospective logistic regression analysis in which G is the dependent variable, and E and D are the independent variables; and where E is constrained to have a null odds ratio association with G in the controls. It is important to note that while the approach is similar to the case-only method, in the sense that the regression analysis also conditions on disease status, both cases and controls contribute to the analysis, in contrast with a case-only analysis in which controls do not contribute. Furthermore, by constraining the main effect of E to be null in the regression, we explicitly encode in the retrospective logistic model, the assumption that E and G are independent in the controls, which, under the rare disease assumption is essentially equivalent to the assumption of gene-environment independence in the underlying population. Finally, by the well known invariance property of odds ratios, the parameter for the main effect of D in the retrospective regression, can be interpreted as the logarithm of the disease odds ratio associated with G among unexposed individuals. In the same vein, the parameter of the interaction between D and E can be interpreted as the logarithm of the $G \times E$ interaction on the odds ratio scale for disease. In fact, as we show below in Section 2, in this simple setting where G and E are both binary, the constrained retrospective MLE of the interaction parameter reduces to the case-only estimator of interaction, and the constrained retrospective MLE of the main effect of D reduces to the all-controls estimator of Modan et al (2001) for the effect of G on D in the unexposed. Therefore, the constrained retrospective logistic regression approach unifies these two methods into a single analytic regression framework, which may be used to formally test for an overall genetic association with disease status. In Section 3, we briefly demonstrate how constrained retrospective regression can accommodate additional covariates L and an exposure variable of a polytomous or continuous nature; we also briefly consider how various forms of genetic coding may be accommodated. It should be noted that when exposure is continuous, the retrospective regression approach allows for the dose-response relationship between exposure and disease risk among non-carriers, to remain completely unrestricted. This latter property is particularly appealing, as it allows for inferences that are completely robust to mis-specification of a model for the odds-ratio association between exposure and the disease outcome; a property, which is not generally shared by standard prospective logistic regression, nor by profile likelihood. In Section 4, we study the performance of the proposed methodology in empirical settings via an extensive simulation

study in which the constrained retrospective MLE is compared to standard prospective logistic regression analysis, as well as to the profile likelihood method. Finally, in Section 5, the methods are illustrated with data from an Israeli study of the interaction between reproductive risk factors and BRCA1/2 in their effects on the risk of ovarian cancer. Some closing remarks are provided in Section 5.

The main appeal of the proposed retrospective regression approach is that (i) it is easy to implement with standard software, (ii) it readily accounts for multiple environmental exposures of a polytomous or of a continuous nature, and also allows for any additional observed covariates, (iii) it is more efficient than ordinary logistic regression (iv) unlike ordinary logistic regression, and the profile likelihood approach of Chatterjee and Carroll (2005), our approach does not require a model for the odds ratio association of the environmental exposure with the disease outcome; as a result it is agnostic and thus completely robust to possible model mis-specification which can happen with an exposure of a polytomous or of a continuous nature.

2 The retrospective regression approach

Binary genetic and environmental factors

In this section, to facilitate the exposition, we introduce our approach in the simple setting of binary genetic and environmental factors. We suppose that the observed data is from an unmatched case-control study in which cases are known to arise from the underlying population according to the saturated logistic regression model

$$\text{logit Pr } (D=1|G, E; \beta) = \beta_0 + \beta_G G + \beta_{GE} GE + \beta_E E \quad (1)$$

where $\beta = (\beta_0, \beta_G, \beta_{GE}, \beta_E)$. Our primary goal is to estimate the overall effects of G , by estimating β_G and β_{GE} . The constrained retrospective logistic regression approach entails fitting by maximum likelihood, the retrospective logistic regression model:

$$\text{logit Pr } (G=1|E, D; \theta) = \theta_0 + \beta_G D + \beta_{GE} ED + \tau E \quad (2)$$

where τ is constrained to be null

$$\tau = \log \left(\text{or}_{GE|D} (1, 1|D=0) \right) = 0,$$

with

$$\text{or}_{AB|C} (a, b|c) = \frac{\text{Pr } (A=a|B=b, C=c) \text{Pr } (A=0|B=0, C=c)}{\text{Pr } (A=a|B=0, C=c) \text{Pr } (A=0|B=b, C=c)}$$

for a and $b \in \{0, 1\}$, defining the odds ratio relating A to B within levels of C , and $\theta = (\theta_0, \beta_G, \beta_{GE})$. Note that the parameters β_G and β_{GE} are shared between models (1) and (2), to reflect the following crucial property of odds ratios:

$$\beta_G = \log \text{or}_{DG|E}(1, 1|E=0) = \log \text{or}_{GD|E}(1, 1|E=0)$$

and

$$\beta_{GE} = \log \frac{\text{or}_{DG|E}(1, 1|E=1)}{\text{or}_{DG|E}(1, 1|E=0)} = \log \frac{\text{or}_{GD|E}(1, 1|E=1)}{\text{or}_{GD|E}(1, 1|E=0)}$$

Furthermore, it should be noted that a null value of τ is perfectly compatible with a non-null value of β_E , since the assumption of gene-environment independence in the controls does not necessarily imply that D is independent of E in non-carriers. Let n_{egd} denote the number of observed subjects with $E = e$, $G = g$ and $D = d$. Also let $n_{\cdot gd} = n_{0gd} + n_{1gd}$. As in Umbach and Weinberg (1994), it is easy to verify that the constrained MLEs of β_G and β_{GE} are

$$\hat{\beta}_G = \log \frac{n_{011}n_{\cdot 00}}{n_{001}n_{\cdot 10}}$$

$$\hat{\beta}_{GE} = \log \frac{n_{111}n_{001}}{n_{101}n_{011}}$$

Next, we illustrate the efficiency advantage of the constrained MLE approach by comparing it to the standard prospective logistic regression approach. To do this, we consider the large sample variance of the constrained MLE $\hat{\beta}_1$ of the effect of G in the unexposed, and of

$$\hat{\beta}_1 + \hat{\beta}_2 = \log \frac{n_{111}n_{\cdot 00}}{n_{101}n_{\cdot 10}}$$

the constrained MLE of the effect of G in the exposed. Recall that the prospective logistic regression MLEs are given by:

$$\tilde{\beta}_G = \log \frac{n_{011}n_{000}}{n_{001}n_{010}}$$

$$\tilde{\beta}_G + \tilde{\beta}_{GE} = \log \frac{n_{111}n_{100}}{n_{101}n_{110}} \quad (3)$$

The estimated variances of $\hat{\beta}_G$ and $\tilde{\beta}_G$ satisfy:

$$\widehat{Var}(\hat{\beta}_G) = \frac{1}{n_{011}} + \frac{1}{n_{001}} + \frac{1}{n_{.00}} + \frac{1}{n_{.10}} \leq \frac{1}{n_{011}} + \frac{1}{n_{001}} + \frac{1}{n_{000}} + \frac{1}{n_{010}} = \widehat{Var}(\tilde{\beta}_G)$$

since $n_{000} = n_{00}$ and $n_{010} = n_{10}$. Similarly, the estimated variances of $\hat{\beta}_G + \hat{\beta}_{GE}$ and $\tilde{\beta}_G + \tilde{\beta}_{GE}$ satisfy:

$$\widehat{Var}(\hat{\beta}_G + \hat{\beta}_{GE}) = \frac{1}{n_{111}} + \frac{1}{n_{101}} + \frac{1}{n_{.00}} + \frac{1}{n_{.10}} \leq \frac{1}{n_{111}} + \frac{1}{n_{101}} + \frac{1}{n_{100}} + \frac{1}{n_{110}} = \widehat{Var}(\tilde{\beta}_G + \tilde{\beta}_{GE})$$

Thus, we can conclude that constrained maximum likelihood estimation yields more efficient estimates of genetic effects than standard prospective regression.

To test for an overall genetic association with the disease outcome, we propose to use the constrained likelihood ratio test

$$2 \left\{ \log L_c(\hat{\theta}) - \log L_c(\hat{\theta}_{null}) \right\} \quad (4)$$

where $\hat{\theta} = (\hat{\theta}_0, \hat{\beta}_G, \hat{\beta}_{GE})$ maximizes the constrained likelihood

$$L_c(\theta) = \prod_i Pr(G_i | E_i, D_i; \theta)$$

and $\hat{\theta}_{null} = (\hat{\theta}_0, \hat{\beta}_G, \hat{\beta}_{GE})$ maximizes the likelihood further constrained by the null hypothesis. In this simple case of binary variables G and E , the likelihood ratio statistic has an asymptotic chi-squared distribution with 2 degrees of freedom under the null. Alternatively, a Wald test statistic may be constructed using the MLE of θ and the corresponding inverse information matrix estimate of its asymptotic covariance matrix. Likelihood ratio and Wald test statistics for other related null hypotheses that may be of interest in this model are easily obtained in a similar fashion.

Polytomous or continuous exposure

Here we consider a straightforward generalization to allow for an environmental exposure of a polytomous or continuous nature, and for the presence of additional covariates L such that G and E are conditionally independent within levels of L but not necessarily so upon marginalizing over any component of L . We suppose that $\omega(E)$ is a user-specified coding for the environmental factor; for example, when E is continuous, $\omega(E)$ may either be equal to a single linear term E , or may incorporate an additional nonlinear term, say $\omega(E) = (E, E^2)^T$. Our generalization entails fitting by maximum likelihood, the constrained retrospective generalized logit model:

$$\text{logit Pr}(G=1|E, D; \theta) = \theta_g + \beta_G D + \beta_{GE}^T \omega(E) D + \theta'_L L \quad (5)$$

where the main effect of exposure representing the conditional log-odds ratio association between G and E given L in the controls

$$\log \left\{ \text{or}_{GE|D,L} (1, E|D=0, L) \right\} \approx \log \left\{ \text{or}_{GE|L} (1, E|L) \right\}$$

is constrained by the independence assumption to be equal to zero. As before,

$$\beta_G = \log \text{or}_{DG|E,L} (1, 1|E=0, L)$$

and the term $\beta_{GE}^T \omega(E)$ now corresponds to the logarithm of the interaction function

$$\log \frac{\text{or}_{DG|E,L} (1, 1|E, L)}{\text{or}_{DG|E,L} (1, 1|E=0, L)}$$

Thus, likelihood ratio statistics to test for non-null genetic effects or interaction parameters may be obtained as in the previous section.

Polytomous genetic factor

We briefly consider a polytomous genetic factor G which we propose to model by adopting a constrained retrospective generalized logistic regression approach, which extends the generalized logistic case-only estimator recently considered by Tchetgen Tchetgen and Robins(2010) and Clarke and Morris (2010):

$$\log \frac{\text{Pr}(G=k|E, D; \theta)}{\text{Pr}(G=0|E, D; \theta)} = \theta_k + \beta_{G,k} D + \beta_{GE,k}^T \omega(E) D + \theta'_{L,k} L$$

where $k = 0, 1, \dots, K$ corresponds to the levels of G . For instance, in the context of GWAS data, $k = 0, 1, 2$ typically corresponds to the three ordered levels for SNP genotype data. As before, the parameters $\{\beta_{G,k}, \beta_{GE,k} : k\}$ of this model can be shown to correspond to the main genetic effects and interaction parameters with exposure on the odds ratio scale. Specific submodels of (??) are obtained by imposing additional constraints on parameters of the model. To illustrate, an additive genotype coding of SNP data is recovered by constraining $\beta_{G,k} = k \times \beta_G$ and $\beta_{GE,k} = k \times \beta_{GE}$ to be linear in k . Tests of non-null genetic effects or interactions are obtained as before via standard likelihood ratio statistics and Wald-type statistics.

3 Robustness of tests of genetic effects

In this section, we consider the robustness against particular forms of model misspecification, of the following methods: (i) standard prospective logistic regression, (ii) profile likelihood and (iii) retrospective logistic regression.

Under the assumption that the disease is rare in the underlying population, profile likelihood estimation (ii) simplifies to maximizing the logarithm of the likelihood

$$\prod_i \frac{\text{odds}_{D|G,E,L}(G_i, E_i, L_i)^{D_i} \Pr\{G_i|L_i\}}{\sum_{d,g} \text{odds}_{D|G,E,L}(g, E_i, L_i)^d \Pr\{G=g|L_i\}}$$

with respect to the unknown parameters of a model for the odds of D given (G, E, L) :

$$\text{odds}_{D|G,E,L}(G, E, L) = \Pr\{D=1|G, E, L\} / \Pr\{D=0|G, E, L\}$$

and of a model for the population density of G , $\Pr\{G|L\}$ (Chatterjee and Carroll (2005)).

Robustness against a mis-specified exposure-disease association

Suppose that G and E are marginally independent in the population of interest and that population stratification is absent, so that it is not necessary to adjust for any variable L in the model. Again, we note that by modeling the genetic variant as the outcome in the regression and disease status as an independent variable, the retrospective approach (iii) does not require a model for the main effect of exposure on disease status and therefore, it is robust to model misspecification of the latter. In contrast, prospective logistic regression and profile likelihood both require that the analyst specify a model for the prospective risk of disease and therefore for the association between E and D , such as that given in equation (1), where it is assumed that $\log \text{or}_{DE|G}(1, E|G=0) = \beta_E E$ is linear in E . When E is continuous, this parametric model may not hold in practice, for instance if the dose-response relationship $\log \text{or}_{DE|G}(1, E|G=0)$ is nonlinear in E . The possibility of a mis-specified model for the effect of exposure raises a concern that the corresponding inferences on genetic effects might be incorrect. However, Tchetgen Tchetgen and Kraft (2010) recently showed that under the null hypothesis of no genetic effects, methods (i) and (ii) respectively yield an MLE of (β_G, β_{GE}) in model (1) that is consistent for the truth, i.e. that converge (in probability) to the correct null value of zero, even when the main effect of exposure $\log \text{or}_{DE|G}(1, E|G=0) = \beta_E E$ is incorrectly specified. Tchetgen Tchetgen and Kraft (2010) formally showed how one can exploit this robustness property, to recover valid inferences on genetic effects in spite of model misspecification. In fact, they showed that valid Wald-type test statistics and confidence intervals for genetics effects can be constructed from the MLE of (β_G, β_{GE}) , provided that a consistent estimator of their variance, say Hubers sandwich estimator of variance is used. For example, the following test statistic may be used:

$$[\tilde{\beta}_G, \tilde{\beta}_{GE}] \hat{\Sigma}^{-1} [\tilde{\beta}_G, \tilde{\beta}_{GE}]^T \sim \chi^2_2 \text{ under the null of no genetic effects}$$

where $(\tilde{\beta}_G, \tilde{\beta}_{GE})$ is the MLE obtained by either method (i) or (ii) and $\hat{\Sigma}$ is the estimated covariance of $(\tilde{\beta}_G, \tilde{\beta}_{GE})$ obtained by extracting the corresponding sub-matrix of the sandwich estimator of the covariance matrix:

$$\left\{ \sum_i \dot{S}_i \right\}^{-1} \sum_i S_i S_i^T \left\{ \sum_i \dot{S}_i^T \right\}^{-1}$$

where S_i and $\dot{S}_i$ are respectively i th individual's estimated contribution to the score equation and Hessian matrix with respect to all unknown parameters in methods (i) or (ii). In contrast, as noted in Tchetgen Tchetgen and Kraft (2010), a test of the null hypothesis of no genetic effects, say via a likelihood ratio test statistic, or a Wald test statistic which uses the inverse of the information matrix as an estimator of the MLE's covariance matrix, will in general have incorrect type 1 error rate, and thus cannot be used in practice to obtain reliable inferences on genetic effects. This is mainly due to the fact that, in contrast to Huber's sandwich covariance estimator, the standard covariance matrix estimator given by the inverse of the information matrix is in general a biased estimator of the covariance matrix of $(\tilde{\beta}_G, \tilde{\beta}_{GE})$ when the model is mis-specified (Tchetgen Tchetgen and Kraft, 2010).

Related results can be shown to hold for methods (i) and (ii) in the presence of categorical covariates L , provided the models are saturated in L under the null. However, when as will often happen in practice, L is multivariate or includes continuous components, saturated models (with respect to L) may not be feasible. Unfortunately, in such settings, as we later illustrate in a simulation study mis-specification of the effect of exposure generally leads to a biased MLE of (β_G, β_{GE}) when either method (i) or (ii) is used, even if the covariates L are correctly modeled. This bias in turn, can result in incorrect inferences of genetic effects. In contrast, as previously noted, a key advantage of estimating (β_G, β_{GE}) by method (iii) is that the retrospective regression approach does not require a parametric model for the effect of exposure on the risk of disease. Therefore, whereas standard logistic regression and the profile likelihood approach may make incorrect inferences when the effect of exposure is mis-specified, an analysis based on the retrospective regression approach is completely immune to this form of model mis-specification whether or not the null hypothesis of no genetic effects holds.

Robustness to violation of the rare-disease assumption

Methods (i) and (ii) equally apply whether or not the disease is rare in the underlying population, although implementation of method (ii) simplifies considerably under the rare disease assumption (see Chatterjee and Carroll, 2005). Thus far, we have only provided justification for using method (iii) when the disease is rare in the underlying population. In this section, we show that this assumption can be relaxed when the goal is to test for genetic

effects; mainly, we show that under the null hypothesis of no genetic effects, the retrospective logistic regression likelihood ratio test (4) retains the correct chi-squared distribution, even when the rare disease assumption does not hold. To prove this claim, it suffices to note that:

$$\begin{aligned} & \text{or}_{GE|D,L}(g, e|D=0, L) \\ &= \frac{\Pr(G=g|E=e, D=0, L) \Pr(G=0|E=0, D=0, L)}{\Pr(G=g|E=0, D=0, L) \Pr(G=0|E=e, D=0, L)} \\ &= \frac{\Pr(D=0|E=e, G=g, L) \Pr(D=0|E=0, G=0, L)}{\Pr(D=0|E=0, G=g, L) \Pr(D=0|E=e, G=0, L)} \\ &\times \frac{\Pr(G=g|E=e, L) \Pr(G=0|E=0, L)}{\Pr(G=g|E=0, L) \Pr(G=0|E=e, L)} = 1 \end{aligned}$$

by the assumption of gene-environment independence, which implies that $\Pr(G=g|E=e, L) = \Pr(G=g|L)$, and by the null hypothesis of no genetic effects which implies that $\Pr(D=0|E=e, G=g, L) = \Pr(D=0|E=e, L)$. Thus, under the independence assumption and the null hypothesis, the log-odds ratio function $\text{or}_{GE|D,L}(1, e|D=0, L)$ between E and G within levels of L of unaffected individuals, is guaranteed to be null as correctly assumed by the constrained retrospective regression approach, whether or not the disease is rare in the underlying population. This in turn implies the previous claim that the corresponding likelihood ratio statistic to test the null hypothesis of no genetic effects or any other related test statistic, is guaranteed to have the correct type 1 error level whether or not the disease is rare in the underlying population, which serves as a basis for recommending the retrospective approach for broader use.

4 Simulation study

We consider the relative performance of the standard logistic regression analysis, the profile likelihood approach and the retrospective regression method under the assumption of conditional gene-environment independence given a covariate L . In our simulation design, L is a discrete variable with three levels $\{0, 1, 2\}$, and probability mass function $\Pr\{L=0\} = 0.6$ and $\Pr\{L=1\} = 0.2$ in the underlying population. G is taken to be binary and two settings are considered: (A) a rare genetic variant setting where $\Pr(G=1|L) = 0.047 \times I\{L=0\} + 0.075 \times I\{L=1\} + 0.120 \times I\{L=2\}$ and (B) a common genetic variant setting where $\Pr(G=1|L) = 0.270 \times I\{L=0\} + 0.380 \times I\{L=1\} + 0.5 \times I\{L=2\}$. We generated the environmental covariate as $E = \min(10, X)$ where X follows the log-normal distribution for which the mean and variance of the underlying normal distribution are $0.7 \times L$ and 1. Given values of (L, G, E) , we generated a binary disease outcome D from the logistic model

$$\text{logit } \Pr(D=1|G, E, L; \beta) = \beta_0 + \beta_G G + \beta_{GE} GE + \beta_E E + \beta_{E^2} E^2 + \beta_L L \quad (6)$$

with $(\beta_0, \beta_E, \beta_L) = (-4, 0.5, 0.1)$ and (β_G, β_{GE}) taking the values given in Tables 1–4 corresponding to four different scenarios. In the first two scenarios (I) and (II), there are no

genetic effects and the association of the environmental exposure is restricted to be linear in scenario (I) i.e. $\beta_G = \beta_{GE} = \beta_{E^2} = 0$, but includes a quadratic term in scenario (II) with $\beta_{E^2} = -0.8$. Settings (III) and (IV) defer from (I) and (II) respectively, only with respect to a non-null main genetic effect and a non-null gene×environment interaction. In each scenario, we generated 4×500 simulated samples, each of the first 500 consisting of 300 cases and 300 controls generated from the above model by sampling cases and controls from a larger random sample of subjects in a rare variant and a common variant setting respectively; and the remaining samples consisting of 500 cases and 500 controls generated from the same model.

Tables 1 through 4 summarize the simulation results pertaining to the regression parameters of primary interest β_G and β_{GE} for scenarios (I) through (IV). Based on these results, we make the following key observations. First, in accordance with theory, the three methods under consideration which recall are (i) prospective logistic regression, (ii) profile likelihood and (iii) retrospective regression, provide essentially unbiased estimates of (β_G, β_{GE}) in the absence of model mis-specification, both under the null model (scenario I) and the non-null model (scenario III). Furthermore, the three methods have correct type 1 error rates under the null and when all models are correctly specified (scenario I). Secondly, a comparison of Monte Carlo variances of the three methods in scenarios (I) and (III) confirms that when model mis-specification is absent, the efficiency gains of methods (ii) and (iii) that exploit the gene-environment conditional independence assumption can be substantial. In fact, the results for (ii) confirm results previously reported by Chatterjee and Carroll (2005) in a similar simulation study. The current simulation study further reveals that while the retrospective regression method (iii) does not model the effect of the continuous exposure factor, the approach which imposes the gene-environment conditional independence, can yield substantial efficiency gains in estimating genetic main and interaction effects similar to those of the profile likelihood approach (see Tables 1 and 3). Thirdly, when in scenarios (II) and (IV), we mis-specify the main effect of the environmental factor on disease risk, both the standard logistic regression approach and the profile likelihood method perform inadequately, with simulation results indicating that the resulting point estimates are largely biased. Furthermore, the 2 degree of freedom tests of genetic effects for methods (i) and (ii) appear to be largely conservative under a mis-specified model (see Table 2) and under-powered to detect non-null effects (see Table 4). In contrast, as we have argued in the previous section, the performance of the retrospective approach remains unaffected by this type of model mis-specification, and its finite sample bias and variance remain optimal across a scenarios considered.

Here we note that although model mis-specification led to conservative inferences based on methods (i) and (ii) in our specific simulation study (Table 2), we have no reason to believe that this will generally be the case, and expect that under the null, mis-specification could equally lead to anti-conservative inferences, reflected by inflated type 1 error rates.

5 Data application

In this section, we illustrate the various methods in an analysis of data from a population-based case-control study based on all ovarian cancer patients identified in Israel between 1

March 1994 and 30 June 1999 (Modan et al. 2001). Two controls per case were selected from the central population registry matching on age within two years, area of birth and place and length of residence. Blood samples were collected on both cases and controls and used to test for the presence of mutation in two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Additional data was collected on reproductive and gynecological history such as parity, number of years of oral contraceptive use and gynecological surgery. The main objective of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecological risk factors of ovarian cancer. To test for interactions between reproductive risk factors and BRCA1/2 in their effects on the risk of ovarian cancer, the authors performed the unadjusted case-only analysis of interaction of Piegorsch et al. (1994) under an assumption that genetic variants and environment factor are unconditionally independent in the population. Chatterjee and Carroll (2005) re-analyzed these data using a fully parametric logistic regression model for disease given the gene, environment and confounding factors L , under the additional conditional independence assumption of gene and environment given the measured covariates L . Tchetgen Tchetgen and Robins (2010) recently implemented a semiparametric case-only estimator of $G \times E$ interaction which adjusts for L in a model for the genetic variant in the cases, but possesses appealing robustness properties in the form of a partial protection against model misspecification of the model for the association of L and G among unexposed cases. However, the results of Tchetgen Tchetgen and Robins (2010) and Chatterjee and Carroll (2005) may not be comparable for two reasons; first, because BRCA1/2 have high penetrance for the risk of ovarian cancer, the disease may not be rare within all levels of G , E and L , as a result, the interaction parameter on the logistic scale may differ from that of the risk ratio scale. Second, even were the disease rare within all strata of G , E and L , Chatterjee and Carroll (2005) assume a fully parametric model for the disease outcome regression model which may result in biased estimates of interactions if their specified working models for the main effect of G (which may include interactions with L), the main effect of E (which may include interactions with L), or the main effect of L are incorrect. As a consequence, the estimator of the interaction parameter of Tchetgen Tchetgen and Robins (2010) is not directly comparable to that obtained by Chatterjee and Carroll (2005). In our re-analysis, we illustrate the retrospective regression method developed in this paper, and thus as in Chatterjee and Carroll (2005), we hypothesize that gene-environment interactions operate on a logistic scale in the underlying population. Specifically, we used data on 832 cases and 747 controls who did not have bilateral oophorectomy, were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. Our primary aim was to test for genetic effects in the presence of interactions between the dichotomous variable representing a person's BRCA1/2 mutation status and her use of oral contraceptives and parity. To illustrate the method with both binary and non-binary environmental exposures, oral contraceptive (OC) use was coded as use for over six years vs use for six years or less, while parity was a count of live births with 10 or more births coded as 10. Thus we fit the regression model

$$\text{logit Pr} \{I(BRCA1/2) = 1\} = \beta_0^\dagger + \gamma(BRCA1/2, OC, Parity) \times D + \beta_L^\dagger L$$

where $\beta_0^\dagger$ is the case-control intercept, and

$$\begin{aligned} & \gamma(BRCA1/2, OC, Parity) \\ &= \beta_{BRCA1/2} + \beta_{BRCA1/2*OC} I(OC > 6) \\ &+ \beta_{BRCA1/2*Parity} Parity + \beta_{BRCA1/2*Parity \geq 2} Parity I(Parity \geq 2) \\ &= \log \left\{ \text{odds}_{DG|E,L} (1, 1|E, L) \right\} \end{aligned}$$

with $G = I(BRCA1/2 = 1)$, $E = (OC, Parity)$.

We assume conditional independence of gene and environmental factors given L , consisting of age (categorical defined by decades), ethnic background (Ashkenazi or non-Ashkenazi), the presence of personal history of breast cancer, a history of gynecological surgery, and family history of breast or ovarian cancer (no cancer vs one breast cancer in the family vs one ovarian cancer or two or more breast cancer cases in the family). The parameter $\beta_{BRCA1/2*Parity > 1}$ encodes the effect of a change point at the value of $Parity = 1$, by modeling the interaction between BRCA1/2 and Parity with a linear spline in $Parity$ with a single knot at 1. We apply the standard logistic regression approach using data both on cases and controls to estimate the interaction between the genetic variant and OC use/parity factors in a regression model which also included main effects for BRCA1/2 indicator, OC use indicator, parity, and L .

$$\begin{aligned} & \text{logit Pr} \{D=1|G, E, L\} \\ &= \beta_0^* + \beta_{OC} I(OC > 6) \\ &+ \beta_{Parity} Parity + \beta_{BRCA1/2} I(BRCA1/2) \\ &+ \gamma(BRCA1/2, OC, Parity) \\ &+ \beta_L' L = \log \left\{ \text{odds}_{D|G,E,L} (G, E, L) \right\} \quad (7) \end{aligned}$$

We also report results for the profile likelihood model with disease risk model given in equation (7) and the following parametric model for specification of carrier frequencies

$$\begin{aligned} & \text{logit Pr} \{G=1|L\} \\ &= \alpha_0 + \alpha_1 I(Age \geq 50) \\ &+ \alpha_2 I(\text{Non-Ashkenazi}) \\ &+ \alpha_3 I(PHBO=1) \\ &+ \alpha_4 I(PHBO=2) \end{aligned}$$

Table 5 summarizes analyses for methods (i) through (iii). All estimates indicate a strong genetic effect among childless women for whom parity=0 and contraceptive use ≤ 6 years. Among these women, BRCA1/2 mutation is strongly associated with ovarian cancer,

producing a remarkable increase in risk of ovarian cancer with odds ratio equal to $\exp(2.7) \approx 15$. The results also indicate a strong statistical interaction between BRCA1/2 and oral contraceptive use and parity. In fact, after adjusting for background variables, the well known protective effects of oral contraceptive use and parity on ovarian cancer risk among BRCA1/2 non-carriers (Modan et al, 2001) appears no longer to generally apply to BRCA1/2 carriers. The results for prospective logistic regression are not statistically significant for any of the interaction parameters, although, point estimates of interactions are comparable (at least within standard errors) for all three methods suggesting that serious model mis-specification may be absent. Methods (ii) or (iii) provide more efficient and comparable estimates of genetic effects when compared to the first method, and also provide more definitive evidence (in terms of statistical significance) of a statistical interaction between BRCA1/2 mutation and parity and oral contraceptive use respectively. According to the results, the elevated risk associated with a mutation in the BRCA1/2 genes appears to increase dramatically with the first birth (with associated odds ratio equal to $\exp(2.68+0.34) \approx 20.5$), and subsequently increases by a lower amount with each additional birth beyond the first birth (with a woman with two children having an associated odds ratio equal to $\exp(2.68+0.34 \cdot 2-0.52) \approx 17.5$). Similarly, the effect of BRCA1/2 increases dramatically in women with more than 6 years of OC use (with associated odds ratio equal to $\exp(2.68+0.97) \approx 38.5$). Methods (i) and (iii) can be obtained from any standard software for performing logistic regression. We used Proc LOGISTIC in SAS 9.1.3 in this paper. The profile likelihood approach was implemented with Proc NLMIXED in SAS 9.1.3. The data were made available by Nilanjan Chatterjee at the Biostatistics Branch of the National Cancer Institute.

6 Conclusion

Case-control studies often have relatively low power for detecting an interaction of small or moderate size between a genetic variant and an environmental exposure and for studying other related hypotheses using the standard prospective logistic regression analysis. When genetic and environmental factors are known to be unrelated within strata of observed covariates, an adjusted case-only approach provides more accurate detection of non-null gene $\times$ environment interaction but does not allow for efficient estimation of genetic effects (Tchetgen Tchetgen and Robins, 2010). The profile likelihood approach uses both cases and controls to efficiently estimate all of the parameters in the disease risk model of interest, by explicitly taking advantage of the independence assumption in a model for the conditional density of the genetic variant. Unfortunately, when the environmental factor is continuous, model mis-specification of the main effect of E on disease risk may have detrimental consequences for the size and the power of tests of genetic effects based on the profile likelihood approach. In this paper, we have considered estimation and tests of genetic effects by modeling both cases and controls in a novel retrospective regression approach. An advantage of the method is that it does not rely on the specification of a model for the main effect of E on D and therefore, it is more robust than available methods.

However, neither the profile likelihood approach nor the proposed retrospective regression approach are immune to bias due to violation of the independence assumption (Albert et al., 2001, Chatterjee and Carroll, 2005). Such bias may still be present even after explicitly

accounting for observable factors L that are known to be related to both G and E , if additional existing common correlates of G and E are not observed and therefore not accounted for.

The ease of implementation of the retrospective approach makes it particularly appealing for use in the analysis of GWAS data where the (conditional) independence assumption is likely to hold for most SNPs. The fact that the method is appropriate for performing tests of genetic effects even when the disease is not rare in the underlying population further adds to its appeal. We should also note that whereas the present paper has focused primarily on the detection and evaluation of genetic effects, the methods may be adapted to a setting where environmental effects are of primary interest. This is achieved by simply substituting E for G as the outcome in the retrospective regression approach. For a continuous exposure variable, the approach would require in addition to the odds ratio model relating the exposure and disease within levels of G and L , a parametric model for the conditional density of exposure given G and L in the controls. In recent work, Tchetgen Tchetgen and Robins (2010), Clarke and Morris (2010) and Tchetgen Tchetgen (2010) have considered a case-only approach for testing and estimating gene-environment interactions which models the density of exposure in cases only, but a fully retrospective approach similar to the one developed in the current paper still remains to be undertaken.

References

- Albert P, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the Case-only Design for Identifying Gene-Environment Interactions. *American Journal of Epidemiology*. 2001; 154(8):687–693. [PubMed: 11590080]
- Breslow NE. Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*. 1996; 91:14–28. [PubMed: 12155399]
- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.
- Cornelis M, Tchetgen Tchetgen E, Liang L, Chatterjee N, Hu F, Kraft P. Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. 2010 Submitted.
- Clarke G, Morris A. A Comparison of Sample Size and Power in Case-Only Association Studies of Gene-Environment Interaction. *American Journal of Epidemiology*. 2010; 171(4):498–505. [PubMed: 20047976]
- Huber, PJ. 1967 Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; The behavior of maximum likelihood estimates under nonstandard conditions; p. 221–223.
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect disease susceptibility loci. *Human Heredity*. 2007 Feb 2; 63(2):111–119. [PubMed: 17283440]
- Modan MD, Hartge P, et al. Parity, oral contraceptives and the risk of ovarian cancer among carriers and no carriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine*. 2001; 345:235–40. [PubMed: 11474660]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*. 1994; 13:153–162. [PubMed: 8122051]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.

- Tchetgen Tchetgen E, Robins J. The semiparametric case-only estimator. *Biometrics*. in Press, Published online March 10 2010.
- Tchetgen Tchetgen E. On the interpretation, robustness and power of varieties of case-only tests of gene-environment interaction. 2010 Submitted.
- Tchetgen Tchetgen E, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is mis-specified. 2010 Submitted.
- Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*. 1997; 16(15):1731–43. [PubMed: 9265696]

Table 1

Simulation results for null genetic effects $\beta_G = \beta_{GE} = 0$ & a correct specification of the main effect of E

		Prospective		Regression		Profile		Lik.		Retrospective		Regression	
		β_G		β_{GE}		β_G		β_{GE}		β_G		β_{GE}	
G rare n=600	bias	0.047		-0.022		-0.037		0.008		-0.038		0.008	
	variance	0.520		0.097		0.244		0.043		0.246		0.043	
	Power [#]	-		0.034		-		0.042		-		0.04	
G rare n=1000	bias	0.064		-0.026		0.006		-0.004		0.006		-0.004	
	variance	0.308		0.061		0.145		0.269		0.145		0.269	
	Power	-		0.042		-		0.048		-		0.048	
G common n=600	bias	0.002		-0.003		-0.007		0.0007		-0.006		0.0007	
	variance	0.144		0.027		0.069		0.014		0.071		0.014	
	Power	-		0.046		-		0.046		-		0.050	
G common n=1000	bias	0.039		-0.013		0.023		-0.006		0.022		-0.005	
	variance	0.079		0.016		0.046		0.009		0.047		0.009	
	Coverage	-		0.040		-		0.054		-		0.056	

[#]Power of Wald-type test statistics of the joint null hypothesis $\beta_G = \beta_{GE} = 0$ based on the inverse information matrix estimator of covariance

Table 2

Simulation results for null genetic effects $\beta_G = \beta_{GE} = 0$ & an incorrect specification of the main effect of E

		Prospective	Regression	Profile	Lik.	Retrospective	Regression
		β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}
G rare $n=600$	bias	0.249	-0.308	-0.064	-0.123	-0.012	0.044
	variance	0.699	0.704	0.200	0.268	0.263	0.432
	Power [#]	-	0.016	-	0.034	-	0.046
G rare $n=1000$	bias	0.361	-0.303	0.066	-0.087	-0.022	-0.031
	variance	0.613	0.380	0.124	0.133	0.167	0.217
	Power	-	0.028	-	0.024	-	0.054
G common $n=600$	bias	0.123	-0.132	0.034	-0.052	-0.002	-0.0005
	variance	0.230	0.138	0.0343	0.060	0.082	0.096
	Power	-	0.020	-	0.022	-	0.032
G common $n=1000$	bias	0.118	-0.130	0.023	-0.047	-0.013	0.003
	variance	0.139	0.080	0.038	0.041	0.049	0.065
	Power	-	0.036	-	0.042	-	0.046

[#]Power of Wald-type test statistics of the joint null hypothesis $\beta_G = \beta_{GE} = 0$ based on the inverse information matrix estimator of covariance

Table 3

Simulation results for nonnull genetic effects $\beta_G = 0.1$, $\beta_{GE} = 0.3$ & a correct specification of the main effect of E

	Prospective		Regression		Profile		Lik.		Retrospective		Regression	
	β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}
G rare n=600	bias*	0.060	-0.021	-0.009	-0.009	-0.0008	-0.008	-0.0008	-0.008	-0.0008	-0.0009	-0.0009
	variance*	0.600	0.098	0.222	0.222	0.029	0.225	0.029	0.225	0.029	0.029	0.029
	Power#	-	0.822	-	0.888	-	-	0.886	-	0.886	-	0.886
G rare n=1000	bias	0.013	-0.001	0.016	0.016	-0.006	0.016	-0.006	0.016	-0.006	-0.006	-0.006
	variance	0.310	0.048	0.118	0.118	0.016	0.118	0.016	0.118	0.016	0.016	0.016
	Power	-	0.976	-	0.988	-	-	0.986	-	0.986	-	0.986
G common n=600	bias	0.002	-0.005	0.003	-0.005	0.0008	-0.004	-0.004	0.0008	-0.004	-0.004	-0.004
	variance	0.202	0.036	0.094	0.015	0.094	0.015	0.094	0.015	0.094	0.015	0.015
	Power	-	0.982	-	0.996	-	-	0.996	-	0.996	-	0.996
G common n=1000	bias	0.009	-0.0006	0.019	-0.005	0.017	-0.004	-0.004	0.017	-0.004	-0.004	-0.004
	variance	0.130	0.023	0.062	0.010	0.063	0.010	0.063	0.010	0.063	0.010	0.010
	Power	-	1	-	1	-	-	1	-	1	-	1

#Power of Wald-type test statistics of the joint null hypothesis $\beta_G = \beta_{GE} = 0$ based on the inverse information matrix estimator of covariance

Table 4

Simulation results for nonnull genetic effects $\beta_G = 0.1$, $\beta_{GE} = 0.3$ & an incorrect specification of the main effect of E

		Prospective	Regression	Profile	Lik.	Retrospective	Regression
		β_G	β_{GE}	β_G	β_{GE}	β_G	β_{GE}
G rare n=600	bias*	0.644	-0.504	0.112	-0.129	0.015	-0.011
	variance*	1.196	0.535	0.137	0.103	0.188	0.137
	Power [#]	-	0.100	-	0.244	-	0.292
G rare n=1000	bias	0.456	-0.382	0.078	-0.114	-0.016	0.004
	variance	0.649	0.255	0.094	0.063	0.130	0.110
	Power	-	0.172	-	0.334	-	0.410
G common n=600	bias	0.307	-0.273	0.082	-0.096	-0.008	0.020
	variance	0.233	0.108	0.057	0.056	0.076	0.089
	Power	-	0.294	-	0.448	-	0.524
G common n=1000	bias	0.324	-0.296	0.078	-0.109	-0.012	0.006
	variance	0.149	0.068	0.043	0.036	0.058	0.059
	Power	-	0.460	-	0.668	-	0.724

[#]Power of Wald-type test statistics of the joint null hypothesis $\beta_G = \beta_{GE} = 0$ based on the inverse information matrix estimator of covariance

Table 5

Parameter estimates, and tests for the genetic effects of BRCA1/2 and interactions between BRCA1/2 and oral contraceptive use and Parity.

	(i)Prosp. Reg.			(ii)Prof. Lik.			(iii)Retro. Reg.		
	Estimate	s.e.	P-value*	Estimate	s.e.	P-value*	Estimate	s.e.	P-value*
$\beta_{BRCA1/2}$	2.70	0.97	0.005	2.68	0.38	0.002	2.61	0.38	<0.0001
$\beta_{BRCA1/2*OC}$	0.56	1.16	0.631	0.90	0.40	0.027	0.97	0.43	0.023
$\beta_{BRCA1/2*Par}$	0.41	0.55	0.459	0.34	0.14	0.018	0.37	0.14	0.010
$\beta_{BRCA1/2*Par^2}$	-0.74	0.70	0.290	-0.52	0.18	0.005	-0.52	0.18	0.005

* The 4df test of an overall genetic effect was highly significant with p-value <0.0001