Estimating Risks and Relative Risks in Case-Base Studies under the Assumptions of Gene-Environment Independence and Hardy-Weinberg Equilibrium

Tina Tsz-Ting Chui¹, Wen-Chung Lee^{1,2*}

1 Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan, **2** Research Center for Genes, Environment and Human Health, College of Public Health, National Taiwan University, Taipei, Taiwan

Abstract

Many diseases result from the interactions between genes and the environment. An efficient method has been proposed for a case-control study to estimate the genetic and environmental main effects and their interactions, which exploits the assumptions of gene-environment independence and Hardy-Weinberg equilibrium. To estimate the absolute and relative risks, one needs to resort to an alternative design: the case-base study. In this paper, the authors show how to analyze a case-base study under the above dual assumptions. This approach is based on a conditional logistic regression of case-counterfactual controls matched data. It can be easily fitted with readily available statistical packages. When the dual assumptions are met, the method is approximately unbiased and has adequate coverage probabilities for confidence intervals. It also results in smaller variances and shorter confidence intervals as compared with a previous method for a case-base study which imposes neither assumption.

Citation: Chui TT-T, Lee W-C (2014) Estimating Risks and Relative Risks in Case-Base Studies under the Assumptions of Gene-Environment Independence and Hardy-Weinberg Equilibrium. PLoS ONE 9(8): e105398. doi:10.1371/journal.pone.0105398

Editor: Balraj Mittal, Sanjay Gandhi Medical Institute, India

Received: March 18, 2014; **Accepted:** July 23, 2014; **Published:** August 19, 2014

Copyright: © 2014 Chui, Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This paper is partly supported by grants from National Science Council, Taiwan (NSC 102-2628-B-002-036-MY3) and National Taiwan University, Taiwan (NTU-CESRP-102R7622-8). No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: wenchung@ntu.edu.tw

Introduction

Many diseases result from the interactions between genes and the environment [1,2]. A case-control study will provide estimates of the genetic and environmental main effects (in terms of odds ratios), and their interactions (in terms of ratios of odds ratios). It is often reasonable to assume that among the non-diseased subjects in the study population, the gene under study is in Hardy-Weinberg equilibrium (which will be achieved within one generation of random mating in any population [3]) and is independent of exposures (genes being constitutional and environmental exposures being exogenous, are often uncorrelated to each other). Previous studies have demonstrated that imposing the dual assumptions of gene-environment independence and Hardy-Weinberg equilibrium can greatly improve the statistical efficiency of a case-control study [4–8].

However, in addition to odds ratios, we may also be interested in knowing the relative and absolute risks of subjects with different genetic and environmental profiles in the population. The relative risk is the ratio of the disease risk for individuals with one specific genetic and environmental profile, to the disease risk for those at a reference level. While indices of relative risk and odds ratio are equally suitable for etiologic inferences, a relative risk (a ratio of two risks) is easier to follow than an odds ratio (a ratio of two ‘odds’; but what is an odds?). The risk itself (or the absolute risk, to be precise) is also important; it is the disease probability for an

individual with a specific genetic and environmental profile, and should be a clinically valuable index. Unfortunately, a case-control design does not provide estimates for the absolute risks; without a rare-disease assumption, estimates for the relative risks are also not provided.

A case-base design is an attractive alternative to the case-control design [9–13]. In contrast to the case-control study which samples the non-diseased subjects in the study base as the control group, the case-base study samples the entire study base without regard to disease status. The design directly produces a relative risk estimate without resorting to the rare-disease assumption [9–12]. Recently, Chui and Lee [13] described a logistic model for case-base study which can be easily fitted using existing statistical software to produce odds ratio estimates and, upon one additional step of simple calculations of the model parameters, relative and absolute risk estimates as well. However, Chui and Lee [13] did not elaborate on how to incorporate the assumptions of gene-environment independence and Hardy-Weinberg equilibrium into the model to further improve statistical efficiency.

In this paper, we show how to analyze case-base study assuming the above dual assumptions. We perform a Monte-Carlo simulation to investigate the statistical performance of the proposed method.

Methods

Case-Base Study Assuming Gene-Environment Independence and Hardy-Weinberg Equilibrium

Let $G = 0, 1$ and 2 represent the number of the variant allele(s) a subject carries. We define two dummy variables: G_1 and G_2 , with $G_1 = 1$, if $G = 1$, and 0 if otherwise; $G_2 = 1$, if $G = 2$, and 0 if otherwise. Let E be the exposure status of a subject which can be in any measurement scale: binary, ordinal or continuous. Let D represent the disease status of a subject, with $D = 1$ for diseased and $D = 0$ for non-diseased. We assume that the disease risk in the study population follows a logistic model:

$$\log \left[\frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)} \right] = \mu + \alpha_1 G_1 + \alpha_2 G_2 + \beta E + \gamma_1 G_1 E + \gamma_2 G_2 E. \quad (1)$$

The $\exp(\mu)$ is the baseline disease odds in the population; $\exp(\alpha_1)$ is the odds ratio of disease for those with $G = 1$ vs. those with $G = 0$; and $\exp(\alpha_2)$, the odds ratio for those with $G = 2$ vs. those with $G = 0$. The $\exp(\beta)$ is the odds ratio associated with the environmental exposure. The $\exp(\gamma_1)$ and $\exp(\gamma_2)$ are the odds ratios associated with gene-environment interactions.

In a case-base study, researchers implement two sampling schemes: the ‘case’ and the ‘control’ sampling schemes [9–13]. The case sampling scheme targets the diseased subjects. Let $S_1 = 1$ indicate that a diseased subject is recruited in the case sample of a case-base study; $S_1 = 0$, otherwise. In such a case sampling scheme:

$$\Pr(S_1 = 1|D, G, E) = \Pr(S_1 = 1|D) = \phi_1 D, \quad (2)$$

where ϕ_1 is a constant between 0 and 1 . Note that the sampling probability depends only on D , not on G and E . In this sampling scheme, the diseased subjects have a constant non-zero probability of being recruited $[\Pr(S_1 = 1|D = 1, G, E) = \Pr(S_1 = 1|D = 1) = \phi_1]$, whereas the non-diseased subjects have a probability of zero of being recruited $[\Pr(S_1 = 1|D = 0, G, E) = \Pr(S_1 = 1|D = 0) = 0]$.

The control sampling scheme targets the entire population (the study base) without regard to disease status. Let $S_0 = 1$ indicate that a subject is recruited in the control sample; $S_0 = 0$, otherwise. Such a control sampling scheme is noted as:

$$\Pr(S_0 = 1|D, G, E) = \phi_0, \quad (3)$$

where ϕ_0 is a constant between 0 and 1 . Note that this sampling scheme essentially is a random sampling of the study population at large; it depends on neither D , G nor E .

The two sampling schemes are assumed to be independent of each other. A diseased subject can be recruited in a case-base study simultaneously in the case sample and in the control sample. The probability of a diseased subject entering the case-base study as a duplicate sample ($S_0 = S_1 = 1$) is:

$$\Pr(S_0 = S_1 = 1|D = 1, G, E) = \Pr(S_0 = 1|D = 1) \times \Pr(S_1 = 1|D = 1) = \phi_0 \phi_1, \quad (4)$$

which depends on neither G nor E . The event of $S_0 + S_1 \geq 1$ indicates that a subject is recruited in a case-base study through case sampling, control sampling or both. Let π be the probability

that a diseased subject recruited in a case-base study can be found in the control sample, i.e.:

$$\begin{aligned} \pi &= \Pr(S_0 = 1|D = 1, G, E, S_0 + S_1 \geq 1) \\ &= \frac{\Pr(S_0 = 1, S_0 + S_1 \geq 1|D = 1, G, E)}{\Pr(S_0 + S_1 \geq 1|D = 1, G, E)} \\ &= \frac{\Pr(S_0 = 1|D = 1, G, E)}{\Pr(S_0 = 1|D = 1, G, E) + \Pr(S_1 = 1|D = 1, G, E) - \Pr(S_0 = S_1 = 1|D = 1, G, E)} \\ &= \frac{\phi_0}{\phi_0 + (1 - \phi_0)\phi_1}. \end{aligned} \quad (5)$$

Again, this depends on neither G nor E .

The maximum likelihood estimate of π and its variance (see Chui and Lee [13]) are:

$$\hat{\pi} = \frac{n_D^{\text{CN}}}{n_D} \quad (6)$$

and

$$\text{Var}(\hat{\pi}) = \frac{\hat{\pi}(1 - \hat{\pi})}{n_D}, \quad (7)$$

where n_D is the total number of *distinct* diseased subject recruited in the case-base study, and n_D^{CN} is the number of diseased subjects recruited in the control sample. Chui and Lee [13] showed that the disease risk in a case-base sample also follows a logistic model:

$$\begin{aligned} \log \left[\frac{\Pr(D = 1|G, E, S_0 + S_1 \geq 1)}{\Pr(D = 0|G, E, S_0 + S_1 \geq 1)} \right] &= \mu^* + \alpha_1 G_1 + \alpha_2 G_2 + \beta E + \gamma_1 G_1 E + \gamma_2 G_2 E, \end{aligned} \quad (8)$$

where $\mu^* = \mu - \log \pi$.

As in Lee et al. [8], we assume that among the non-diseased subjects in the study population, the gene (G) is independent of environmental exposure (E) [the first equality in the following Equation (9)] and in the Hardy-Weinberg equilibrium (the second equality), i.e.:

$$\begin{aligned} \Pr(G|E, D = 0) &= \Pr(G|D = 0) \\ &= (1 - p)^2 \times 2^{G_1} \times \left(\frac{p}{1 - p} \right)^{G_2} \\ &= \Pr(G = 0|D = 0) \times \exp(G_1 \log 2 + \delta G), \end{aligned} \quad (9)$$

where p is the allele frequency and $\delta = \log\left(\frac{p}{1-p}\right)$, the log allele frequency odds, among the non-diseased subjects in the study population. Combining Equations (8) and (9), the likelihood function for a case-base study under the assumptions of gene-environment independence and Hardy-Weinberg equilibrium is found to be (Exhibit S1):

$$\begin{aligned} \Pr(D, G|E, S_0 + S_1 \geq 1) &= \frac{\exp(G_1 \log 2 + \delta G + \mu^* D + \alpha_1 G_1 D + \alpha_2 G_2 D + \beta E D + \gamma_1 G_1 E D + \gamma_2 G_2 E D)}{\sum_{d=0}^1 \sum_{g=0}^2 \exp(g_1 \log 2 + \delta g + \mu^* d + \alpha_1 g_1 d + \alpha_2 g_2 d + \beta E d + \gamma_1 g_1 E d + \gamma_2 g_2 E d)}. \end{aligned} \quad (10)$$

Model (10) above has exactly the same form of the likelihood function of a 1:5 matched case-control study. [The denominator in Model (10) has a total of six terms, corresponding to one ‘case’ and

five ‘controls’.] Therefore, one can adopt the ‘counterfactual approach’ [8] to fit Model (10). To be precise, one first creates a recruitment indicator, $R=1$, for each and every distinct subject actually recruited in the case-base study. Next, one creates a total of five counterfactual subjects (all of them with $R=0$) to each recruited subject; the exposure status (E) of the five counterfactual subjects is deliberately set to be exactly the same as the recruited subject to whom they are matched, but the disease status (D) and gene (G) are different. (The five subjects represent the five different ways of making $[D, G]$ -different counterfactuals.) Treating R as the outcome variable, one then performs a conditional logistic regression analysis (using existing statistical software, such as SAS) with the following regression equation: $G_1 \log 2 + \delta G + \mu^* D + \alpha_1 G_1 D + \alpha_2 G_2 D + \beta ED + \gamma_1 G_1 ED + \gamma_2 G_2 ED$, based on the above created 1 (factual): 5 (counterfactuals) matched data. The results are the conditional maximum likelihood estimates of the total 7 parameters in Model (10), together with their variance-covariance matrix, Σ , a 7×7 matrix.

Among the parameter estimates obtained from a fitting of Model (10) to data, $\hat{\delta}$ is an estimate for log allele frequency odds in Equation (9), $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are estimates for log genetic odds ratios in Model (1), $\hat{\beta}$ is an estimate for log environmental odds ratio in Model (1), and $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimates associated with gene-environment interactions in Model (1). The $\hat{\mu}^*$ estimate from Model (10) is to be further combined with the $\hat{\pi}$ estimate from Equation (6) to provide an estimate for the log baseline disease odds in Model (1):

$$\hat{\mu} = \log(\hat{\pi}) + \hat{\mu}^* \quad (11)$$

Because $\hat{\pi}$ is independent of $\hat{\mu}^*$ [13], the variance of $\hat{\mu}$ is:

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}[\log(\hat{\pi}) + \hat{\mu}^*] \\ &= \text{Var}[\log(\hat{\pi})] + \text{Var}(\hat{\mu}^*). \\ &= \frac{1}{\hat{\pi}^2} \times \text{Var}(\hat{\pi}) + \text{Var}(\hat{\mu}^*) \quad (12) \\ &= \frac{1 - \hat{\pi}}{n_D^{\text{CN}}} + \text{Var}(\hat{\mu}^*). \end{aligned}$$

Now, we can estimate the absolute and relative risks. An estimate of the (absolute) disease risk for subjects in the study population with ($G_1 = g_1, G_2 = g_2, E = e$) is:

$$\begin{aligned} \widehat{\text{risk}}_u &= \frac{\exp(\hat{\mu} + \hat{\alpha}_1 g_1 + \hat{\alpha}_2 g_2 + \hat{\beta} e + \hat{\gamma}_1 g_1 e + \hat{\gamma}_2 g_2 e)}{1 + \exp(\hat{\mu} + \hat{\alpha}_1 g_1 + \hat{\alpha}_2 g_2 + \hat{\beta} e + \hat{\gamma}_1 g_1 e + \hat{\gamma}_2 g_2 e)} \\ &= \frac{\exp[\log(\hat{\pi}) + \hat{\mu}^* + \hat{\alpha}_1 g_1 + \hat{\alpha}_2 g_2 + \hat{\beta} e + \hat{\gamma}_1 g_1 e + \hat{\gamma}_2 g_2 e]}{1 + \exp[\log(\hat{\pi}) + \hat{\mu}^* + \hat{\alpha}_1 g_1 + \hat{\alpha}_2 g_2 + \hat{\beta} e + \hat{\gamma}_1 g_1 e + \hat{\gamma}_2 g_2 e]} \quad (13) \\ &= \frac{\exp[\log(\hat{\pi}) + \mathbf{u}\hat{\eta}]}{1 + \exp[\log(\hat{\pi}) + \mathbf{u}\hat{\eta}]}, \end{aligned}$$

where $\mathbf{u} = [1 \ g_1 \ g_2 \ e \ g_1 e \ g_2 e]$ is a 1×6 gene-environment profile vector and $\hat{\eta} = [\hat{\mu}^* \ \hat{\alpha}_1 \ \hat{\alpha}_2 \ \hat{\beta} \ \hat{\gamma}_1 \ \hat{\gamma}_2]^t$ is a 6×1

vector of parameter estimates. Because $\hat{\pi}$ is independent of all the other parameters [13], the variance of the estimate (in logit scale) is:

$$\begin{aligned} \text{Var}[\text{logit}(\widehat{\text{risk}}_u)] &= \text{Var}[\log(\hat{\pi}) + \mathbf{u}\hat{\eta}] \\ &= \text{Var}[\log(\hat{\pi})] + \text{Var}(\mathbf{u}\hat{\eta}) \quad (14) \\ &= \frac{(1 - \hat{\pi})}{n_D^{\text{CN}}} + \mathbf{u}\mathbf{\Omega}\mathbf{u}^t, \end{aligned}$$

where $\mathbf{\Omega} = \text{Var}(\hat{\eta})$ is readily available by simply deleting the first row and the first column of Σ . [$\hat{\delta}$ from Model (10) plays no role in the disease risk estimation.] Let the reference group be those subjects in the study population with ($G_1 = g_1^*, G_2 = g_2^*, E = e^*$), and let $\mathbf{u}^* = [1 \ g_1^* \ g_2^* \ e^* \ g_1^* e^* \ g_2^* e^*]$ be the gene-environment profile vector for them. An estimate of the relative risk is therefore:

$$\begin{aligned} \widehat{\text{RR}}_{u/u^*} &= \widehat{\text{risk}}_u / \widehat{\text{risk}}_{u^*} \\ &= \frac{\exp[\log(\hat{\pi}) + \mathbf{u}\hat{\eta}]}{1 + \exp[\log(\hat{\pi}) + \mathbf{u}\hat{\eta}]} \bigg/ \frac{\exp[\log(\hat{\pi}) + \mathbf{u}^*\hat{\eta}]}{1 + \exp[\log(\hat{\pi}) + \mathbf{u}^*\hat{\eta}]} \quad (15) \end{aligned}$$

Using the delta method, the variance of the estimate (in log scale) is:

$$\text{Var}[\log(\widehat{\text{RR}}_{u/u^*})] = (\widehat{\text{risk}}_u - \widehat{\text{risk}}_{u^*})^2 \times \left(\frac{1 - \hat{\pi}}{n_D^{\text{CN}}} + \mathbf{w}\mathbf{\Omega}\mathbf{w}^t \right), \quad (16)$$

$$\text{where } \mathbf{w} = \frac{(\widehat{\text{risk}}_u - 1)\mathbf{u} - (\widehat{\text{risk}}_{u^*} - 1)\mathbf{u}^*}{\widehat{\text{risk}}_u - \widehat{\text{risk}}_{u^*}}.$$

Monte-Carlo Simulations

For simplicity, we assume a binary exposure E ($E=0, 1$) and a biallelic gene with genotype G ($G=0, 1, 2$). For the non-diseased subjects in the study population, we assume gene-environment independence and Hardy-Weinberg equilibrium, with the exposure prevalence (for $E=1$) and the allele frequency of the variant allele both being set at 0.5. The disease probabilities of subjects in the study population are assumed to follow the logistic model [Model (1)]. We assume an autosomal recessive gene with a genetic odds ratio of 2.0 [$\alpha_1=0$ and $\alpha_2=\log(2)$], an environmental odds ratio of 2.5 [$\beta=\log(2.5)$] and a gene-environmental interaction odds ratio of 2.0 [$\gamma_1=0$ and $\gamma_2=\log(2)$]. The disease prevalence in the study population is set at 0.1. Therefore, the six disease risks are: $\text{risk}_{G=0,E=0}=0.0441$, $\text{risk}_{G=1,E=0}=0.0441$, $\text{risk}_{G=2,E=0}=0.0845$, $\text{risk}_{G=0,E=1}=0.1034$, $\text{risk}_{G=1,E=1}=0.1034$, and $\text{risk}_{G=2,E=1}=0.3157$, respectively. And the five relative risks are (with $G=0, E=0$ as the reference level) $\text{RR}_{G=1,E=0}=1.0000$, $\text{RR}_{G=2,E=0}=1.9155$, $\text{RR}_{G=0,E=1}=2.3449$, $\text{RR}_{G=1,E=1}=2.3449$, and $\text{RR}_{G=2,E=1}=7.1587$, respectively. A case-base study is conducted in the study population (population size: 100,000) with a case sampling probability (ϕ_1) of 0.05 and a control sampling probability (ϕ_0) of 0.005. Under such a sampling scheme, the case-base study is expected to recruit a total of 500 distinct diseased and 500 distinct non-diseased subjects. Exhibit S2 presents the SAS code for simulating data.

Using the formulas derived by Cheng and Lin [6], we compare the relative efficiencies of the estimates in the conditional logistic regression [with the assumptions of gene-environment indepen-

Table 1. Simulation results for a biallelic gene and a binary exposure with and without the assumptions of gene-environment independence and Hardy-Weinberg equilibrium.

Log relative risk or logit absolute risk	Estimate		Variance		Coverage probability of 95% confidence interval		Average length of 95% confidence interval	
	True value	with assumptions	without assumptions	with assumptions	with assumptions	without assumptions	with assumptions	without assumptions
$\log RR_{G=1, E=0}$	0.0000	0.0135	0.0069	0.0533	0.9534	0.9535	0.8980	1.0537
$\log RR_{G=2, E=0}$	0.6500	0.6632	0.6579	0.0616	0.9506	0.9495	0.9608	1.0275
$\log RR_{G=0, E=1}$	0.8522	0.8524	0.8602	0.0830	0.9540	0.9513	1.1204	1.4368
$\log RR_{G=1, E=1}$	0.8522	0.8663	0.8629	0.0621	0.9533	0.9518	0.9740	1.1097
$\log RR_{G=2, E=1}$	1.9683	1.9846	1.9810	0.0611	0.9513	0.9517	0.9619	1.0288
$\logit(risk_{G=0, E=0})$	-3.0762	-3.1022	-3.0980	0.0695	0.9530	0.9531	1.0258	1.0984
$\logit(risk_{G=1, E=0})$	-3.0762	-3.0888	-3.0914	0.0348	0.9495	0.9500	0.7284	0.7608
$\logit(risk_{G=2, E=0})$	-2.3831	-2.3959	-2.3934	0.0330	0.9485	0.9484	0.7096	0.7351
$\logit(risk_{G=0, E=1})$	-2.1599	-2.1850	-2.1707	0.0807	0.9515	0.9501	1.0914	1.3479
$\logit(risk_{G=1, E=1})$	-2.1599	-2.1714	-2.1720	0.0411	0.9542	0.9520	0.7937	0.8772
$\logit(risk_{G=2, E=1})$	-0.7736	-0.7828	-0.7765	0.0343	0.9506	0.9552	0.7258	0.8414

doi:10.1371/journal.pone.0105398.t001

dence and Hardy-Weinberg equilibrium, Model (10)] relative to the corresponding estimates in Chui and Lee's [13] unconditional logistic regression [without the dual assumptions, Model (8)]. [Relative efficiency of A method relative to B method is defined as the ratio of the variance of B estimator and that of A estimator. A larger-than-one relative efficiency implies a better statistical performance (larger power and better precision, etc) for A method as compared to B method.]

In regard to the estimates for relative risks and absolute risks, we perform a total of 10,000 simulations to compare the performances of the two approaches (the conditional logistic regression with the dual assumptions vs. the unconditional logistic regression without the dual assumptions). The means of the estimates for relative risks (in log scale) and absolute risks (in logit scale) are calculated. The variance of an estimate is calculated as the sample variance of the estimates. We also calculate the coverage probabilities and the average lengths of the 95% confidence intervals for the estimates. [The coverage probability of a 95% confidence interval for a parameter estimate is the probability that the interval covers the true value of the parameter in a repeated sampling (simulation) experiment.]

Results

The relative efficiency of the proposed method (with the dual assumptions of gene-environment independence and Hardy-Weinberg equilibrium), as compared to Chui and Lee's method (without the dual assumptions) [13], is shown in Exhibit S3. It can be seen that exploiting the dual assumptions can greatly improve statistical efficiency. The relative efficiencies are 1.30 ($\hat{\alpha}_1$), 1.44 ($\hat{\alpha}_2$), 1.37 ($\hat{\beta}$), 1.57 ($\hat{\gamma}_1$), and 1.85 ($\hat{\gamma}_2$), respectively.

Table 1 shows the simulation results for the estimates for relative and absolute risks. Using either approach, the estimates of relative and absolute risks are approximately unbiased. [The bias of a parameter estimate is the difference between the mean of the parameter estimates in the simulation experiment and the true value of that parameter; compare the column labeled 'Estimate' and the column labeled 'True value' in Table 1.] The 95% confidence intervals also achieve adequate coverage probabilities for both approaches. However, the variances and the average lengths of the confidence intervals using the present method, which imposes the dual assumptions of gene-environment independence and Hardy-Weinberg equilibrium, are much smaller than Chui and Lee's method [13] which imposes neither assumption.

Discussion

Lee et al. [8] previously discussed how to relax the Hardy-Weinberg equilibrium for case-control studies. This can also be applied to the present context of case-base studies. Assuming only the gene-environment independence assumption, the case-base likelihood becomes:

$$\Pr(D, G|E, S_0 + S_1 \geq 1) = \frac{\exp(\varepsilon_1 G_1 + \varepsilon_2 G_2 + \mu^* D + \alpha_1 G_1 D + \alpha_2 G_2 D + \beta ED + \gamma_1 G_1 ED + \gamma_2 G_2 ED)}{\sum_{d=0}^1 \sum_{g=0}^2 \exp(\varepsilon_1 g_1 + \varepsilon_2 g_2 + \mu^* d + \alpha_1 g_1 d + \alpha_2 g_2 d + \beta Ed + \gamma_1 g_1 Ed + \gamma_2 g_2 Ed)},$$

where $\varepsilon_1 = \log[\Pr(G=1|D=0)/\Pr(G=0|D=0)]$ and $\varepsilon_2 = \log[\Pr(G=2|D=0)/\Pr(G=0|D=0)]$ are gene-frequency-related parameters (log genotype frequency odds among the non-diseased subjects in the study population, to be precise). By comparison, the likelihood function (Model 10) where both assumptions are imposed contains only one gene-frequency related parameter (δ).

If the study population is not homogeneous, but is instead composed of a number of population strata, a case-base study is also vulnerable to population stratification biases just as a case-control study can be. Assume that there are a total of q stratum; the case-base likelihood conditioned on stratum indicators, $t_1, \dots, t_q$, is:

$$\Pr(D, G | t_1, \dots, t_q, E, S_0 + S_1 \geq 1) = \frac{\exp(G_1 \log 2 + \sum_{i=1}^q \delta_i t_i G + \sum_{i=1}^q \mu_i^* t_i D + \alpha_1 G_1 D + \alpha_2 G_2 D + \beta ED + \gamma_1 G_1 ED + \gamma_2 G_2 ED)}{\sum_{d=0}^1 \sum_{g=0}^2 \exp(g_1 \log 2 + \sum_{i=1}^q \delta_i t_i g + \sum_{i=1}^q \mu_i^* t_i d + \alpha_1 g_1 d + \alpha_2 g_2 d + \beta Ed + \gamma_1 g_1 Ed + \gamma_2 g_2 Ed)}.$$

The interaction terms $t_i \times G$ ($t_i \times D$) allows the allele frequency odds (the background disease odds) to vary between different population strata. To use this model, one needs to know in advance the stratum to which each and every study subject belongs.

In this paper, we present a method to analyze the case-base study exploiting the assumptions of gene-environment independence and Hardy-Weinberg equilibrium with common statistical packages. When both assumptions are met, the simulation results show that the method is approximately unbiased and has adequate coverage probabilities of the 95% confidence intervals. It also results in smaller variances and shorter confidence intervals as

compared to a previous assumption-free method for a case-base study.

Supporting Information

Exhibit S1 Derivation of the likelihood function for a case-base study under the assumptions of gene-environment independence and Hardy-Weinberg equilibrium. (PDF)

Exhibit S2 SAS code for simulating data. (PDF)

Exhibit S3 Calculation of the relative efficiency of the proposed conditional logistic regression method (with the dual assumptions of gene-environment independence and Hardy-Weinberg equilibrium) as compared to the unconditional logistic regression method (without the dual assumptions). (PDF)

Author Contributions

Conceived and designed the experiments: WCL. Performed the experiments: TTTC. Analyzed the data: TTTC. Contributed reagents/materials/analysis tools: WCL. Contributed to the writing of the manuscript: TTTC WCL.

References

- Hunter DJ (2005) Gene-environment interactions in human diseases. *Nat Rev Genet* 6:287–298.
- Olden K (2007) Commentary: From phenotype, to genotype, to gene-environment interaction and risk for complex diseases. *Int J Epidemiol* 36:18–20.
- Sham P (1998) *Statistics in Human Genetics*, Oxford University Press Inc., New York, pp39–43.
- Umbach DM, Weinberg CR (1997) Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Stat Med* 16:1731–1743.
- Chatterjee M, Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92:399–418.
- Cheng KF, Lin WJ (2005) Retrospective analysis of case-control studies when the population is in Hardy-Weinberg equilibrium. *Stat Med* 24:3289–3310.
- Chen YH, Kao JT (2006) Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies. *BMC Genet* 7:43.
- Lee WC, Wang LY, Cheng KF (2010) An easy-to-implement approach for analyzing case-control and case-only studies assuming gene-environment independence and Hardy-Weinberg equilibrium. *Stat Med* 29:2557–2567.
- Kupper LL, McMichael AJ, Spirtas R (1975) A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Assoc* 70:24–528.
- Miettinen OS (1982) Design options in epidemiologic research: an update. *Scand J Work Environ Health* 8(S1):7–14.
- Sato T (1992) Maximum likelihood estimation of the risk ratio in case-cohort studies. *Biometrics* 48:1215–1221.
- Sato T (1994) Risk ratio estimation in case-cohort studies. *Environ Health Persp* 102(S8):53–56.
- Chui TTT, Lee WC (2013) A regression-based method for estimating risk and relative risk in case-base study. *PLoS ONE* 8:e83275.