SHORT REPORT

Assessing effects of disease genes and gene-environment interactions: the case-spouse design and the counterfactual-control analysis

Wen-Chung Lee, Chin-Hao Chang

J Epidemiol Community Health 2006;**60**:683–685. doi: 10.1136/jech.2005.043554

Background: Assessing joint genetic and environmental contributions to disease risk is the central issue in many genetic epidemiological studies. To characterise the effects of a gene, the case-control study may suffer from the problem of population stratification bias. For a late onset disease, recruiting control subjects into case-parents and case-sibling studies may be difficult.

Methods: Two novel approaches to analysing case-spouse data are introduced: the 1:1 case-counterfactual-control analysis (genotype swapping between the case and their spouse) and the 1:5 case-counterfactual-controls analysis (allele swapping).

Results: Both can be implemented using statistical packages that allow matched analysis (the conditional logistic regression) to yield valid estimates of the genotype relative risk, the gene-environment interaction parameter, the gene-sex interaction parameter, and the gene-environment-sex three factor interaction parameter (if desired), if certain assumptions are fulfilled.

Conclusion: Because of the ease in recruiting subjects, and in collecting and analysing data, this approach makes a convenient tool for gene characterisation.

The occurrences of most human diseases are the result of interplay between genes and environmental factors. Assessing joint genetic and environmental contributions to disease risk is the central issue in many genetic epidemiological studies[1]: Cho *et al*,[2] using a case-control design, found that the presence of the *HOGG1* variant would increase 1.6-fold a subject's risk for nasopharyngeal carcinoma (the genetic effect); van Rooij *et al*,[3] also using a case-control design, found that the presence of the *GSTT1* variant in a smoking mother would increase 3.2-fold her children's risk for non-syndromic oral clefting, while *GSTT1* variant or smoking alone on the part of the mother would not increase her children's risk (the gene-environment interaction). Through studies such as these, the effects of genes on human diseases are better characterised.

However, the case-control study may suffer from the problem of population stratification bias, if it is conducted in a ''stratified population''.[4 5] (The study population is ''stratified'' if it is composed of two or more strata in which both baseline disease risks and genotype frequencies differ across the strata.) Family based association studies,[1 6–8] such as case-parents and case-sibling studies, have become popular in the past decade. The studies recruit the family members (parents and sibling(s), respectively) of the cases as control subjects. Because the case and their ''matched'' control(s) are now from the same family (and thus from the same population stratum as well), the case-parents and the case-sibling studies are robust to population stratification biases. However, these studies may have difficulties in recruiting control subjects. In practice, parents/siblings may live elsewhere and be hard to reach or may refuse to participate. Also, parents may already have died or the case may be the only child in the family.

In a recent paper, Lee[9] described a new approach, the case-spouse study. As the name suggests, the study recruits the spouses as the control subjects for the cases. For a study aiming at a late onset disease, such as non-insulin dependent diabetes, cardiovascular disease, Alzheimer disease, many forms of cancers, etc, recruiting spouses should be much easier than recruiting parents or siblings. (As a norm, an adult will get married and live together with their mate; thus, recruiting a married couple for study should be easy.) Recruiting spouses may even be easier than recruiting unrelated controls. (If a person gets sick, their spouse will often be the one who brings the sick person to medical attention. This suggests that hospitals/clinics may provide a convenient setting for conducting a case-spouse study.)

However, Lee's study[9] is for ''gene mapping''. In this paper, however, we were interested in applying the case-spouse design for ''gene characterisation''—that is, to assess the effects of a disease gene and any possible gene-environment interaction. In particular, we introduce a counterfactual-control approach[10–14] to analysing the case-spouse data.

THE CASE-SPOUSE STUDY WITH COUNTERFACTUAL-CONTROL ANALYSIS

A certain number of case-spouse pairs were recruited. Genotyping was done for the cases and their spouses, but the information on environmental exposures was collected for the cases only. In each and every pair, a ''genotype swap'' was performed between the case and their spouse. (The case and their spouse exchange their genes, two alleles at a time.) The genotype swapped case in a case-spouse pair is then regarded as a counterfactual control for the case. Thus 1:1 case-counterfactual-control data are artificially created. Next ''allele swaps'' between the case and their spouse are performed in each and every pair, first one allele and then two alleles at a time. The resulting five allele swapped cases (four single allele swaps and one genotype swap) in a case-spouse pair are then regarded as five counterfactual controls for the case. And 1:5 case-counterfactual-controls data have thereby been created. Note that the counterfactual control(s) of a case created in this way has the same environmental exposures and sex as the case themself.

The above 1:1 case-counterfactual-control data can be analysed legitimately (as if these counterfactual controls are the real subjects) using statistical packages that permit matched analysis (conditional logistic regression), such as Egret or SAS (proc phreg), if the following assumptions are

met (a proof is in the appendix, which is available from the authors and on line http://www.jech.com/supplemental):

- (A1) Mating is restricted to subjects in the same stratum.
- (A2) Mating is independent of genotype, for males and females in each and every stratum.
- (A3) Environmental exposures are independent of genotype, for males and females in each and every stratum.
- (A4) The genotype frequencies for males are equal to the corresponding frequencies for females, in each and every stratum.

The 1:5 case-counterfactual-controls data can also be analysed legitimately using the standard conditional logistic regression, if the above A1–A4 together with the following A5 and A6 assumptions are met (a proof is in the appendix/ available from the authors):

- (A5) There is no imprinting effect for the gene under study.
- (A6) Each and every stratum is in Hardy-Weinberg equilibrium.

Maximum likelihood estimates of the genotype relative risk, the gene-environment interaction parameter, the gene-sex interaction parameter, and the gene-environment-sex three factor interaction parameter (if desired) can be obtained readily from the computer output. Note that these parameters have ''risk ratio'' interpretations as compared with the usual ''odds ratio'' interpretations, because a log-linear risk model (as compared with the logistic risk model[15]) is invoked. Also readily available are the standard errors, confidence intervals, and p values for these parameters. If desired, the model that fits the data the best can be chosen among the competing models using standard model selection techniques. Note however, the main effects of the ''matching factors'' and the interaction terms between them are non-identifiable. That is, the relative risks for environmental exposure(s), for sex, and the exposure by exposure, exposure by sex interactions cannot be estimated using such a counterfactual control approach to case-spouse data.

COMPARISON WITH OTHER METHODS

Case-spouse study with actual-spouse analysis

An alternative method for analysing the case-spouse data is to treat the spouse as they are. That is, the spouse in a case-spouse pair gets to retain their own gene, sex, and environmental exposures, and serves as a matched control for the case. The 1:1 case-actual-spouse data are then analysed using the ordinary conditional logistic regression.[15] This approach needs only the A1 assumption and can do away with the A2–A6 assumptions entirely. However, the case-actual-spouse approach is contingent on collecting all the risk factors for the disease (for the cases and for the spouses as well) and incorporating them in the conditional logistic regression model. Otherwise, it will (without the A2–A4 assumptions) suffer from confounding biases. The approach is also more expensive in terms of data collection and is less efficient statistically (more parameters have to be estimated, one main effect parameter for each risk factor additionally), as compared with the case-counterfactual-control approach. On the other hand, it should be emphasised again that the validity and the efficiency gain associated with a counterfactual-control analysis are hinged on making more assumptions.

Case-control study with stratum matching

Another possibility is to abandon the case-spouse design altogether (thereby obviating the need for the A1 assumption) and to resort to a case-control study with stratum matching[16] (although the second may be a more complicated design). Here ''stratum matching'' means that for each case, (a number of) control(s) are recruited from the population at large that has the same race, ethnicity, nationality, ancestry, and birthplace, as the case themself. Such stratum matching will lessen the impact of population stratification bias, but generally will not eliminate it completely (the self reported race, ethnicity, nationality, ancestry, and birthplace may not resolve the true stratum membership for each and every subject in a study). It is difficult to predict which design, the case-spouse design or the stratum matched case-control design, will attain smaller bias—the bias for the case-spouse design depends on the degree of inter-strata admixing (marriages) in the population, and the bias for the stratum matched case-control design depends on the matching error associated with the self reported ''stratum delineating variables'' used in a study.

Unmatched case-control study

For the sake of simplicity, one always has the option to perform a conventional case-control study—that is, an unmatched design with both the cases and the controls recruited from the population at large. This approach will suffer from population stratification bias, if the study population is stratified. However, Wacholder *et al*[17] pointed out that ''there will be only a small bias from population stratification in a well designed case-control study of genetic factors that ignores ethnicity among non-Hispanic U.S. Caucasians of European origin.'' In fact, they estimated the bias to be about 1%, and no more than 10% in general. If stratum matching of some sort is involved, either through a real marriage as in the case-spouse design or through an artificial pairing as in the stratum matched case-control design, the bias should be even smaller. But on the other hand, the results of a ''matched'' design apply only to the strata defined by the cases themselves. And thus the external validity of the study (for example, if the cases are from a single hospital) can be questioned.

DISCUSSION

In this study, we assume that the genotype frequencies for males are equal to the corresponding frequencies for females, in each and every stratum. If the gene under study is not at the sex chromosome, the genotype frequencies at conception should be the same for both sexes because of random segregation of sex chromosomes and autosomes. Furthermore, if the gene under study is not in linkage disequilibrium with another gene (or is the gene itself) that affects survival through gestation or over time, the genotype frequencies at different ages should be the same for males and females alike. If the validity of this assumption is a concern, one should check the genotypes of the spouses to see whether the frequencies vary with sex or age.

Environmental exposure and mating are assumed to be independent of genotype for males and females in each and

Policy implications

Because of the ease in recruiting subjects, and in collecting and analysing data, the case-spouse design with counterfactual-control analysis makes a convenient tool for gene characterisation. However, it should be noted that no empirical data exist to refute or confirm the validity of this theoretically elegant approach, and interpretation of gene-environment interaction without a measure of environment main effect is limited in scope.

every stratum. These are reasonable assumptions, unless of course the gene under study, beyond its possible effects on disease susceptibility, is also a gene (or in linkage disequilibrium with a gene) that influences extent of exposure or mating choice. If these assumptions do not hold, the counterfactual-control approach will be biased and lose its advantage in statistical efficiency. Note that mating does not need to be independent of exposure in any given population stratum. ''Exposure'' dependent mating is common in the real world (for example, between subjects of similar physique, religion, socioeconomic background, etc). Assortative mating of this kind by itself does not invalidate the counterfactual-control approach to case-spouse study.

The assumption that mating is restricted to subjects in the same stratum deserves special attention. Marriages between subjects in different strata are of course possible, but such incidents are probably rare in practice. If the admixture proportion of a stratified population is really very high, such a population will mix itself in a few generations. (Weinberg[18] pointed out that ''cross-marriages, which are needed to cause artifactual [sic] asymmetries across spouses, themselves quickly obliterate the stratification.'') Thus, in any reasonably steady state population, the spouse of a case should provide a near ideal genetic control for the case themself. If this is a concern for a particular case-spouse pair, one can genotype a panel of ''single nucleotide polymorphisms'' across the genome for the case and their spouse and test whether they are coming from the same stratum using Lee's method.[19]

The 1:5 case-counterfactual-controls analysis needs two more assumptions than the 1:1 case-counterfactual-control analysis does—that is, (A5) there is no imprinting effect for the gene under study and (A6) each and every stratum is in Hardy-Weinberg equilibrium. Therefore, the 1:5 analysis may have a narrower range of valid application than the 1:1 analysis. However, if A1–A6 assumptions can be met, a 1:5 analysis will yield effect estimates that are more stable (smaller standard errors and narrower confidence intervals) as compared with the 1:1 analysis.

The concept of counterfactual controls[10–14] is the basic tenet of this paper. To a decent geneticist or epidemiologist, the case-spouse design with counterfactual-control analysis must seem to be counterintuitive in many ways. Firstly, spouses are the opposite sex to the cases. Instead of the usual same sex matching in many epidemiological studies, the case-spouse study is, and truly is, ''counter-matched'' on sex. Secondly, some diseases (for example, prostate cancer) occur exclusively in one sex. However, the case-spouse study still uses the spouses as the controls, without regard to the fact that the spouses now (the wives of prostate cancer patients) are not the ''population at risk''. Finally, a gene can exert its effect predominantly in one sex but have none or little effect in the other (perhaps through the interaction with endogenous hormones, the levels of which are quite different between men and women). Accordingly, one may want to conduct a study that recruits subjects, cases, and controls alike, of only one sex, that is, the sex in which the gene is expressed. After all, why bother to recruit subjects of the other sex, if it is already known that they are unaffected by the gene? However, the case-spouse design with counterfactual-control analysis still applies, by including the gene-sex interaction term in the regression model, or alternatively, by recruiting cases of the gene active sex only, together with their (opposite sex) spouses.

Because of the ease in recruiting subjects, and in collecting and analysing data, the case-spouse design with counterfactual-control analysis makes a convenient tool for gene characterisation. However, it should be noted that no empirical data exist to refute or confirm the validity of this theoretically elegant approach, and interpretation of gene-environment interaction without a measure of environment main effect is limited in scope.

The appendix is available from the authors and on line (http://www.jech.com/supplemental)

Authors' affiliations
W-C Lee, C-H Chang, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University

Funding: this study was partly supported by a grant from the National Science Council, Republic of China.

Conflicts of interest: none.

Correspondence to: Dr W-C Lee, Rm 536, No 17, Xuzhou Road, Zhongzheng District, Taipei 100, Taiwan; wenchung@ha.mc.ntu.edu.tw

Accepted for publication 27 January 2006

REFERENCES

1 **Weinberg CR**, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000;**152**:197–203.
2 **Cho EY**, Hildesheim A, Chen CJ, *et al*. Nasopharyngeal carcinoma and genetic polymorphisms of DNA repair enzymes XRCC1 and HOGG1. *Cancer Epidem Biomark Prev* 2003;**12**:1100–4.
3 **van Rooij IA**, Wegerif MJ, Roelofs HM, *et al*. Smoking, genetic polymorphisms in biotransformation enzymes, and nonsyndromic oral clefting: a gene-environment interaction. *Epidemiology* 2001;**12**:502–7.
4 **Ewens WJ**, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Human Genet* 1995;**57**:455–64.
5 **Witte JS**, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999;**149**:693–705.
6 **Spielman RS**, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Human Genet* 1993;**52**:506–16.
7 **Spielman RS**, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Human Genet* 1998;**62**:450–8.
8 **Lee WC**, Chang CH. Estimating genotype relative risks in case-parental control studies: an optimal weighting approach. *Am J Epidemiol* 2000;**152**:487–92.
9 **Lee WC**. Genetic association studies of adult-onset diseases using the case-spouse and case-offspring designs. *Am J Epidemiol* 2003;**158**:1023–32.
10 **Self SG**, Longton G, Kopecky KJ, *et al*. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics* 1991;**47**:53–61.
11 **Maclure M**. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 1991;**133**:144–53.
12 **Farrington CP**, Nash J, Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *Am J Epidemiol* 1996;**143**:1165–73.
13 **Khoury MJ**, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996;**144**:207–13.
14 **Zaffanella LE**, Savitz DA, Greenland S, *et al*. The residential case-specular method to study wire codes, magnetic fields, and disease. *Epidemiology* 1998;**9**:16–20.
15 **Breslow NE**, Day NE. *Statistical methods in cancer research*. Vol 1. *The analysis of case-control study*. Lyon: IARC Scientific Publications, 1980.
16 **Lee WC**. Case-control association studies with matching and genomic controlling. *Genet Epidemiol* 2004;**27**:1–13.
17 **Wacholder S**, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000;**92**:1151–8.
18 **Weinberg CR**. Invited commentary: making the most of genotype asymmetries. *Am J Epidemiol* 2003;**158**:1033–5.
19 **Lee WC**. Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Ann Hum Genet* 2003;**67**:618–19.