

## Epidemiologic and Genetic Approaches in the Study of Gene-Environment Interaction: an Overview of Available Methods

N. Andrieu<sup>1</sup> and A. M. Goldstein<sup>2</sup>

### INTRODUCTION

The interest in conducting studies to examine gene-environment interaction is increasing for most chronic and complex diseases such as cancer. This increased interest is mostly due to considerable advances in molecular genetic techniques. Gene-environment studies are motivated by different situations including: 1) The detection of major genes that do not have estimated lifetime risks that reach 100 percent (i.e., incomplete penetrance); the incomplete penetrance may result from the role of other factors in disease etiology, such as environmental factors, that may or may not interact with the genetic factors. In such a situation, the genetic factor would be considered as the major risk factor and the environmental exposure as a modifier. 2) The identification of gene-environment interaction as a common biologic process in pharmacogenetic studies of complex diseases; pharmacogenetics hypothesizes that hereditary factors that control the metabolism of carcinogens or other toxic substances may modulate risk of disease. Indeed, different genotypes (for example, *GSTM1* null genotype versus non-null) seem to respond differently to environmental risk factors. In this situation, the environmental factor would be considered as the major risk factor and the genetic factor as a modifier. 3) Inconsistent associations across studies between a disease and a suspected risk factor; one reason for the inconsistencies is that relevant risk factors may be difficult to detect, possibly due to heterogeneity in the studied populations. One source of heterogeneity could be unknown genetic susceptibility that predisposes to differential

environmental sensitivity. Indeed, ignored gene-environment interaction could easily conceal effects of an environmental factor on risk of disease. Also, one may observe inconsistent results across studies trying to identify the mode of inheritance of a disease segregating in families. Indeed, ignoring gene-environment interaction, when it exists, may decrease the power in segregation analyses and affect inferences (1).

These situations and others have led to increased study of interaction between genetic and environmental factors. In this presentation we review methods to detect gene-environment interactions according to whether or not a surrogate/actual measure of the genetic factor(s) is(are) available. Interaction is a model-dependent concept and needs to be defined accordingly. For example, a gene-environment interaction exists if the joint effect of the genetic factor and the environmental exposure differs from the product of the risks for the individual factors on a multiplicative scale, and from the sum of the background disease rate and the excess rates for the environmental exposure and for the genetic factor on an additive scale.

### THE GENETIC SUSCEPTIBILITY IS KNOWN AND A SURROGATE OR ACTUAL MEASURE IS AVAILABLE

This circumstance is possible when either the gene itself has been identified or a closely linked marker is available.

#### Case-only study

In case-only studies, the association between an environmental exposure and a genotype is examined among case subjects only (affected subjects with a given disease). This method has been proposed recently to evaluate gene-environment interaction in disease etiology (2, 3), and has been reviewed by Khoury and Flanders (4). Therefore, we will only briefly present this method.

The study design uses a case-series which follows the same epidemiologic principles of case selection as for any case-control study. The exposure effect is

Received for publication November 10, 1997, and accepted for publication August 7, 1998.

Abbreviations: OR, odds ratio; CYP2D6, cytochrome P450 2D6; RR, relative risk; TDT, transmission disequilibrium test.

<sup>1</sup>Unité de Recherche en Épidémiologie des Cancers, Institut de la Santé et de la Recherche Médicale (U351), Institut Gustave-Roussy, Villejuif Cedex, France.

<sup>2</sup>Genetic Epidemiology Branch, National Cancer Institute, Bethesda, MD.

Reprint requests to Nadine Andrieu, Unité de Recherche en Épidémiologie des Cancers, Institut de la Santé et de la Recherche Médicale (U351), Institut Gustave-Roussy, 94805 Villejuif Cedex, France.

assessed only among cases. Cases without the susceptibility genotype form the pseudocontrol group, and cases with the susceptibility genotype form the pseudo-case group. These two groups are compared with respect to the prevalence of the environmental exposure (the risk factor) with the nonexposed pseudocases being the referent group. Odds ratios and confidence intervals are obtained using standard crude analyses or multivariate models to adjust for other covariates. As demonstrated by Piegorsch et al. (2), this approach is valid to estimate interaction between a gene and environmental exposure if the exposure and genetic factor occur independently and the disease is rare. If these assumptions are valid, the odds ratio is the interaction effect (see Smith and Day (5)) as measured in a regular case-control study under a multiplicative model (for more details see Khoury and Flanders (4)).

This study design has been used to detect interaction between environmental factors and genetic markers in tumors. One such example, by Lehrer et al. (6), was a case-only study to assess an association between spontaneous abortion and a polymorphism in the human estrogen receptor gene on 31 women with estrogen-receptor positive breast tumors. The authors found an association between the rarer allele, called B-variant genotype allele, and a history of spontaneous abortion.

Assuming that the environmental exposure and genetic factors occur independently, analyses of cases-only studies offer better precision for estimating gene-environment interactions than those based on both cases and controls (subjects not affected by the disease under study), i.e., there are smaller standard errors due to elimination of control group variability (2). The power for detecting gene-environment interactions in case-only studies is comparable to the power for assessing a main effect in a classic case-control study (7). The advantage of this design is that only cases are needed, thus avoiding the difficult and often unsatisfying selection of appropriate controls. The main disadvantage is that the main effects of the genetic and environmental factors cannot be estimated. In addition, even if many biologically plausible gene-environment interaction models should cause departures from multiplicative effects (8), case-only studies would miss gene-environment models with departures from additivity. For these reasons, Umbach and Weinberg (9) recently proposed an alternative method that keeps the advantages of the case-only design while simultaneously allowing estimation of the main effects. The proposed incomplete-data case-control design collects both genotype and environmental exposure data from the cases but only environmental

exposure or only genotype data from the controls. The estimation of main effect(s) is possible only if the genotype and environmental exposure are independent (i.e., occur independently) and the studied disease is rare. Umbach and Weinberg (9) proposed a maximum likelihood method based on log-linear models which allows imposition of the independence assumption, whereas usual logistic regression does not permit such an imposition. The authors showed that their method may need fewer than half as many individuals as methods that do not impose the gene-environment independence assumption to reach the same power for detecting gene-environment interaction. This approach allows estimation of interaction that is a departure from either multiplicativity or additivity in the relative risk. However, in case-only and incomplete-data case-control designs, the required assumption of independence between a gene and environmental factor may be violated when the environmental exposure and the genetic factor both vary with an unmeasured variable. To ensure the validity of the critical independence assumption, a random sample of controls with both genetic and exposure data should be collected (9).

The sib-pair and affected-pedigree-member methods from classic genetics studies may also use case-only approaches (10–12). The major aim of these methods is to detect genetic linkage to chromosomally localize susceptibility genes involved in the studied disease. These methods examine the number of alleles at a given locus that are inherited identical by descent (in the sib-pair method) or identical by state (in the affected-pedigree-member method) between affected relatives. For example, under the assumption of no genetic linkage, the expected distribution of alleles shared by descent between siblings is 25 percent for zero alleles, 50 percent for one allele, and 25 percent for two alleles. Departures from this distribution suggest linkage between the disease and the studied marker. These studies could also be performed incorporating environmental factors into the analyses by stratifying the sample according to the environmental exposure (e.g., exposed versus nonexposed). A difference in the expected allele sharing by level of exposure could be statistically evaluated and may be due to gene-environment interaction. However, absence of linkage in one or another exposure category may reflect loss of statistical power due to reduced subsample sizes. Further, a difference in allele sharing by exposure level might result from other factors including etiologic heterogeneity. Finally, neither the genetic nor the environmental exposure main effect, nor their interaction effect, may be estimated by these methods.

### Case-control study design using unrelated controls

The case-control study design using unrelated controls is the most commonly used design in studies of gene-environment interaction. Using unexposed subjects with no susceptibility genotype as the referent group, odds ratios for all other groups can be estimated. Adjustment for potential confounding variables may be done by using stratification or multivariate approaches. For this type of study, the meaning of interaction must be specified explicitly, because interaction is a model-dependent concept. The choice of model (e.g., multiplicative or additive) will depend on many factors including the overall goals of the study. Departures from multiplicativity can be assessed by the interaction effect (5) which is the joint odds ratio for the exposure and the genotype divided by the product of the odds ratios (OR) for the effect of exposure alone and of the genotype alone ( $OR_{interaction} = (OR_{exposure, genotype} / OR_{exposure} OR_{genotype})$ ). Departures from an additive model of risk may also be assessed by dividing the joint odds ratios for the exposure and the genotype by the sum of the odd ratios for the effect of exposure alone and of the genotype alone minus 1 ( $OR_{interaction} = (OR_{exposure, genotype} / (OR_{exposure} + OR_{genotype} - 1))$ ). An interaction effect  $>1$  indicates a greater than multiplicative/additive effect between the exposure and the genotype, while an interaction effect  $<1$  indicates a less than multiplicative/additive effect (13).

The power of this traditional case-control study design for assessing interaction was first studied examining interaction between unspecified risk factors (5, 14, 15), and then focusing on gene-environment interaction (16, 17). Among the parameters needed to determine the power to detect a gene-environment interaction, the frequencies of the environmental exposure and the genetic factor appeared to be the most important in determining the number of cases required (16). In fact, results from Hwang et al. (16) and Goldstein et al. (17) suggest that case-control designs may be used to detect gene-environment interaction only when the environmental factor and the genetic factor are common. Indeed, certain combinations of odds ratios and frequencies of the genetic factor and the environmental factor make study sizes prohibitive. For example, a rare gene (gene frequency = 0.01) and a common environmental exposure (exposure prevalence = 0.3), both increasing the risk of disease by 2.0, would require approximately 12,000 cases and 12,000 controls to detect an interaction of 3 between the gene and the environmental factors on a multiplicative scale (with 80 percent power) (17). Higher gene frequencies lead to more reasonable sample sizes. As shown in

Hwang et al. (16), when the frequency of exposure and the frequency of the gene range between 0.30 and 0.70, about 200 cases (and 400 controls) would be required to detect an interaction greater than 4 (with 80 percent power).

Recently, this study design has been used in pharmacogenetic studies to evaluate putative associations of susceptibility genes involved in metabolism of carcinogens and cancers taking into account carcinogen exposure. For example, Bouchardy et al. (18) investigated the interaction between smoking exposure and cytochrome P450 2D6 (CYP2D6) activity, which is involved in debrisoquine metabolism, on the risk of lung cancer. In a case-control study of 128 lung cancers and 157 controls, they found a positive interaction between increasing tobacco consumption and increasing CYP2D6 activity.

This design has also been used to evaluate etiologic heterogeneity according to the presence of a familial factor. In this approach, a positive family history of disease was used as a surrogate for the gene. For example, some studies have been performed on breast cancer looking for interactions between familial and reproductive factors. Results tend to be inconsistent across studies with no clear pattern for the interaction effect (19–23). In these studies, it is likely that, rather than the study design, the surrogate measure of the genetic factor (i.e., family history) is the reason for the inconsistent results. Indeed, family history of breast cancer may be due to the joint effect of environmental factors common to family members and to genetic factors which could differ from family to family. This joint effect could be highly heterogeneous and thus produce a poor surrogate measure of the genetic factor.

The main advantage of this design is that the main effects of the environmental exposure and genetic susceptibility, as well as their interactive effect, may be estimated. The main disadvantage is that this design may not be appropriate for the study of gene-environment interaction involving rare genes or uncommon environmental exposures (assuming moderate values of the interaction effect). In addition, population stratification or genetic admixture might adversely influence this design. The extent of the problem produced by population stratification is still unknown and needs to be evaluated.

In order to increase the power to detect a gene-environment interaction when one of the factors under study is rare, one has to increase the number of cases and controls. However, measurement of the gene and/or environmental factor would likely be too costly to obtain for a large population. One alternative design may be the two-stage or multistage design.

### Two-stage case-control study: counter-matching and balanced design

When one of the factors is rare in a gene-environment interaction study, the two-stage or multistage design may be appropriate. These designs increase the numbers of cases and controls with the rare factor without prohibitively increasing the number of measurements to perform.

For two-stage study designs, samples of cases and controls are drawn from the population at risk. After classification according to an exposure of interest, subsamples of cases and controls are selected for purposes of covariate assessment (for example, assessment of the genetic factor when the environmental factor is the exposure of interest; assessment of the environmental factor when the genetic factor is the exposure of interest). Approaches include balancing the numbers of exposed and unexposed cases and controls as well as counter-matching in which controls are counter-matched, rather than matched, to cases. The goal of both approaches is to improve the statistical efficiency of exposure risk estimates compared with the classic random case-control study (24, 25).

In studies of gene-environment interaction, counter-matching may be considered as an alternative method for increasing efficiency in interaction detection while keeping a reasonable number of cases and controls. Counter-matching has been recently proposed by Langholz and Clayton (25) as a method of sampling controls for nested case-control studies. The goal of counter-matching is to maximize the number of discordant case-control pairs from which information comes in a matched case-control study. Towards that goal, in gene-environment interaction assessment, one scenario for counter-matching would be to use a surrogate of the genetic factor, such as family history. This measure is generally available on the entire cohort/population at risk at stage I, whereas a measure of the gene itself would be too costly to obtain for the whole cohort. Assuming the disease is rare, at stage II all cases from the cohort/population-at-risk would be sampled. Each case's risk set would be stratified by the surrogate of the gene, namely family history of disease, and controls for that risk set would be selected from the strata other than the case's stratum. The higher the sensitivity of the surrogate, namely family history, of the genetic factor, the more case-control pairs discordant for the genetic factor. Thus, the gain in efficiency will depend on how predictive family history of disease is for the gene of interest.

Alternatively, one could sample based on an environmental exposure or on both the environmental factor and the genetic factor. A partial likelihood has been developed to estimate exposure effects in counter-

matching (26) using weighting that takes into account the probability that subjects were selected from specific strata. This method has been shown to increase the efficiency of main effect estimation by approximately 25 percent compared with classic random sampling (27). In gene-environment interaction assessment, counter-matching on both the gene and environmental factor has been shown to be more efficient than a standard nested case-control study or designs which counter-matched on either the environmental factor or the gene. The relative efficiencies were influenced mainly by the frequency of the risk factors of interest (i.e., not the surrogates) and the sensitivity and specificity of the factors' surrogates (B. Langholz and D. Thomas, University of Southern California, Los Angeles, California, unpublished data). Studies to examine additional scenarios and evaluate cost, efficiency, and feasibility need to be conducted to determine the utility of these approaches for interaction assessment.

A more intuitive approach to studying a rare factor would be to oversample for the rare covariate of interest. Two-stage case-control studies using balanced oversampling is such a design. If either the gene or the environmental factor under study is rare, rather than choose a subset at random, cases and controls may be sampled nonrandomly, that is, oversampled for the rare factor (called exposure). Thus, for example, stage II would consist of selecting all cases (if the disease is rare) exposed to the rare factor and sampling an equal number in each of the three other categories. Then the other factor (called covariate) involved in the gene-environment interaction is measured in this subsample. The oversampling is taken into account in the analysis to obtain unbiased estimates of the exposure effect and interaction effect in which the exposure is involved (24, 28). Cain and Breslow (28) investigated the efficiency of a balanced design compared with a random sampling case-control design particularly for estimating exposure-covariate interaction. A rare exposure was assumed with a prevalence of 0.05 and an odds ratio of 2.0. The prevalence of the covariate was set at 0.3 with several odds ratio values and several values of correlation between the exposure and covariate as measured by the odds ratio in the control group. The results showed that the balanced design was always much more efficient than a random sampling design for estimating the exposure-covariate interaction in terms of standard errors for both a rare and a common disease. The efficiency decreased as the correlation between the exposure and covariate increased but still remained substantial. For example, when the degree of confounding between the exposure and covariate was set at 5.0, the relative efficiency of the

balanced design was greater than 3 for a common disease and equal to 2 for a rare disease.

Langholz and Borgan (26) considered a model analogous to the one above (29) for their counter-matching design. They called it "balanced counter-matched design" and assessed the efficiency of their method in estimating exposure-covariate interaction. Instead of choosing a surrogate for the rare exposure of interest, they considered the exposure of interest itself for sampling stratification. Using the same set of parameters as Breslow and Cain (29), they obtained similar efficiencies in interaction estimation.

In both two-stage balanced and balanced-counter-matched designs, the rare exposure is assumed to be known for the entire population at stage I. Thus, if a rare genetic or environmental factor requires expensive investigation to be measured, measurement may not be affordable at the first stage. One could then propose a three-stage study with two successive sampling stages. Stage I would consist of selecting all cases (if the disease is rare) exposed to an inexpensive surrogate variable for the factor under study, and sampling equal numbers of unexposed cases, and exposed and unexposed controls. The factor assessment could then be performed on the stage II subsample. Stage III would consist of selecting all cases exposed to the rare factor and sampling an equal number of unexposed cases, and exposed and unexposed controls. The environmental or genetic factor of interest for the gene-environment interaction evaluation would be collected in this stage III subsample. Such an approach needs to be statistically evaluated before being considered. At present, identifying good surrogates for the factor(s) of interest, and the costs associated with measuring a genetic or environmental factor on large numbers of subjects, may be the major determinants for deciding whether or not to conduct multistage, balanced and counter-matched designs.

#### Case-control study design using related controls

To our knowledge, a case-control study design using related controls has not yet been applied in studies of gene-environment interaction. In this approach, each case is matched to one or more unaffected relative(s) and a standard conditional logistic regression is performed to assess gene-environment interaction. This method does not result in biased estimates and does yield consistent estimates even when there is a correlation in risk factors under study between relatives (30, 31). There was no bias under the hypothesis that the risk factors remained constant over time. However, bias may occur if there is a correlation, within matched case-control pairs, on unmeasured risk factors (32). Bias may also occur if there is an interaction between

an unknown genetic factor or other correlated unmeasured risk factor and the factors under study (33). Witte et al. (34) evaluated different types of relatives for use as controls in gene-environment interaction studies discussing the advantages and disadvantages of each type of relative in terms of overmatching, variation in exposures over time (cohort effect), or differential survival. For gene-environment interaction studies, one recommendation was to restrict relatives to those who are concurrent in age and calendar time with the cases to avoid age and cohort effects. In addition, cousins were suggested as potentially being the best eligible controls, although costs of finding and ascertaining this control group have not been thoroughly examined.

One advantage of this study design is that relatives are easily identified and may be more willing to participate in a research study. In addition, this study design permits estimation of the main effects of environmental factors and genetic susceptibility and their interaction effect. There may also be a gain in efficiency for detecting gene-environment interaction involving a rare gene compared with a traditional case-control study approach because of oversampling on the genetic factor by using related controls. Studies to assess the power of this design are under way.

The use of unaffected siblings has also been proposed in an extension of the sib-pair methodology (described in the case-only study section above) to study the role of environmental factors and specific gene loci, and to provide evidence of gene-environment interaction (35). Cases are affected siblings of the proband (case who led to family selection) and controls are randomly chosen from unaffected siblings of the proband. The genetic factor for each case and control is defined as the number of alleles shared identical by descent with the proband at the studied locus. For example, the odds ratios associated with the genetic factor are obtained by comparing the share-2 alleles and share-1 allele groups with the baseline risk group, share-0 alleles. Covariates, including environmental exposures, may be investigated for main effects and for evidence of interaction with the studied locus by using stratification or multivariate analyses. An alternative matched design has also been proposed for matching on family. This requires families with at least one affected and one unaffected sibling in addition to the proband limiting the number of available families. Power and efficiency for gene-environment interaction detection using these approaches has not yet been investigated.

Khoury and Flanders (4) have proposed another not yet evaluated method for gene-environment interaction study. The case-parental control method, also

known as the transmission disequilibrium test (TDT) method (36), which consists of comparing the genotype of each case with the genotype of a fictitious control comprised of the nontransmitted alleles from each parent, was first proposed for genetic markers-disease association studies. This method has been extended by stratifying case subjects according to the presence, or absence, of an environmental exposure. The odds ratios associated with the genetic factor can be obtained for nonexposed and exposed cases. Differences in the estimates across strata may reflect gene-environment interaction. This study design does not allow for assessing the main effect of the exposure. Also, differences in estimates across strata are not restricted to resulting from only gene-environment interaction. Differentiating between the various reasons for the differences may be difficult.

Maestri et al. (37) used the TDT method to investigate the relation between oral clefts and markers associated with five candidate genes, included covariates such as type of cleft, race, family history, and maternal smoking, to test these covariates as effect modifiers. They detected a significant interaction between maternal smoking and the transmission of markers near two candidate genes (*TGFA* and *TGFB3*, both transforming growth factor).

The advantage common to these case-related control study designs is the elimination of controls whose genetic backgrounds differ systematically from that of cases by using relatives as the controls. Some of these approaches permit estimation of the main effects of the environmental exposures and genetic factors, as well as their interaction effects, whereas other approaches only allow estimation of the interaction effect. One disadvantage of these designs is the potential for overmatching on exposures of interest leading to a loss in efficiency. The extent of this loss will depend on the exposures of interest and their correlation among relatives.

#### **THE GENETIC SUSCEPTIBILITY IS UNKNOWN AND, THUS, A SURROGATE MEASURE IS NOT AVAILABLE**

When genetic susceptibility is unknown, three different kinds of methods may be used to detect gene-environment interaction. The first approach consists of searching for modification in risk associated with exposure involved in gene-environment interaction using different types of controls. The second approach simultaneously assesses genetic susceptibility and gene-environment interaction assuming different models of inheritance as in segregation analyses. The third approach consists of using twin studies where monozygotic twins who have all their genes in common are

compared with dizygotic twins who have half of their genes in common, as is the case for full siblings. Different analytic methods have been proposed for gene-environment assessment.

#### **Case-control study design using relatives of cases and population- or hospital-based controls**

The proposed study design is a case-control study using two types of controls, relatives of cases and unrelated controls. The risk associated with environmental exposure estimated using relatives of cases as the controls is compared with that estimated using population-based or hospital-based unaffected persons as the controls. Indeed, it has been demonstrated that the estimates of odds ratios of environmental exposures may be modified by using relatives as controls compared with population-based controls when risk factors interact with an unknown underlying genetic factor (33, 38).

The relative risks of disease associated with environmental exposure using siblings of cases as controls were always higher than those obtained using population-based controls when the joint effects of genotype and environmental exposure on disease risk were more than multiplicative. The relative risks of disease associated with environmental exposures using relatives of cases as controls were identical to those obtained using population-based controls when the genotype and the environmental exposure had multiplicative joint effects on disease risk (33, 38). That finding had also been demonstrated by Goldstein et al. (30), where they showed the equivalence of the relative risks across control groups if the studied environmental exposure remained constant over time. Additionally, Andrieu and Goldstein (33) showed that the relative risks of disease were identical when the genetic factor had no effect on disease risk within the group exposed to the environmental factor or within the nonexposed group. The difference in the estimates of odds ratios was dependent on the amount of interaction between the genetic and environmental factors and on the genetic correlation between relatives. An interaction with a rare genetic factor produced only a small difference between the environmental factor odds ratios. The difference between the environmental factor odds ratios became larger when the genetic factor was more common. Let us note that an unmeasured environmental factor clustered in families might also lead to differences in magnitude of the studied environmental factor odds ratios if these two factors interact.

Few studies have been performed using two sets of control groups, related and unrelated controls. Two such studies investigated effects of reproductive fac-

tors on breast cancer (39, 40). The differences in the estimated risks using relative-controls (i.e., sister controls) and hospital-based controls were not directly tested for statistical significance, and only descriptive presentations of the difference in the point estimates regarding the risks was proposed. Both studies reported modifications in point estimates of disease risk associated with age at menarche (39) and number of abortions (39, 40).

Neither of the methodological studies mentioned above assessed the power to detect differences in risks (i.e., to detect interactions), so it is difficult to know how informative selecting two types of controls would be. Power assessment of this study design is currently being investigated (N. Andrieu, INSERM, Villejuif, France, and A. M. Goldstein, US National Cancer Institute, Bethesda, Maryland, unpublished data, 1998). One disadvantage of the design is that it does not allow for measuring risk of disease associated with the unknown genetic factor or the genetic factor-environmental exposure interaction, since the genetic factor is unknown. The advantage is that a possible gene-environment interaction may be indicated by a difference in the estimates of exposure risk. This difference in risk is a function of the genetic factor-environmental exposure interaction value. If this study design is shown to be efficient, it could be used to detect relevant risk factors as well as understand their role in the etiology of diseases where unknown genetic factors may interact with studied environmental risk factors.

### Twin study designs

Originally, twin studies were proposed to evaluate either genetic influences on disease comparing risks in monozygotic and dizygotic cotwins of affected cases (called probands) or environmental exposure effects comparing exposed and nonexposed twins in monozygotic cotwin-control studies. Recently, Ottman (41) developed a method of testing gene-environment interaction in twin data ascertained through affected twins and proposed two epidemiologic measures:  $RR_E$ , relative risk of disease in exposed versus unexposed cotwins, stratified by zygosity and proband exposure status, and  $RR_z$ , relative risk of disease in monozygotic versus dizygotic cotwins, stratified by exposure status of proband and cotwin. Differences in  $RR_E$  between monozygotic and dizygotic cotwins or differences in  $RR_z$  between exposed and unexposed cotwins suggest gene-environment interaction. Calculations for assessing power of this method were performed on a hypothetical study of 200 monozygotic and 200 dizygotic pairs. The results showed that the differences in  $RR_z$  and  $RR_E$  were expected to be small

(less than twofold for an interaction value of 5) except under very extreme conditions (interaction value greater than 100). Conversely, if large differences in  $RR_E$  or  $RR_z$  are observed in a twin study, they are likely to reflect strong gene-environment interaction (41).

Another method has been proposed using unselected samples of twins (42) allowing also for gene-environment interaction as well as main effects. However, statistical power for this method has not been studied.

These recent methods in twin studies have the advantage of permitting estimation of the effects of the environmental exposure and genetic factor as well as their interaction effects, but only under very high levels of interaction. Moreover, possible confounding by unmeasured common environmental factors might further decrease the power to detect interactions. Thus, the number of required twin pairs to assure power sufficient to detect interactions of moderate values, and the difficulties of collecting data on twins, might make this type of study prohibitive.

### Family study design and segregation analyses

This proposed study design is a family study with no separate control group. Families are recruited through one or more diseased members and data on disease status and covariates are collected on all family members. The aim of segregation analysis is to identify the most likely mode of inheritance explaining a trait's distribution in families. Generally, the transmission pattern is modeled as a major gene segregating like a dominant, recessive, or codominant trait, and/or a polygene or other types of correlation of phenotypes among family members (which may have either a polygenic basis or unmeasured common environmental factor or both). Earlier models for segregation analyses did not allow direct testing for gene-environment interactions (43–45). Instead, researchers stratified families into different groups based on environmental exposure levels and then analyzed the mode of inheritance of the trait within each group. Differences in inheritance models suggested the possibility of interaction between a genetic susceptibility and the environmental factor (46, 47). This approach was an extension to an approach commonly used to assess genetic heterogeneity in diseases (48, 49) with the idea that one source of heterogeneity might be due to gene-environment interaction. This approach was often difficult or infeasible because of the variability in exposure within families.

More recently, direct gene-environment interaction assessment using segregation analysis approaches has been possible (50–54). These models allow for estimating genetic, environmental, and gene-environment

interaction risks using somewhat different approaches. The regressive models (52) were constructed by specifying a relation between each person's phenotype for a studied trait and a set of explanatory variables including the person's major genotype, the phenotype of relatives to take into account residual family dependencies of unspecified origin (genetic or environmental), and measured covariates. Studied traits were able to be either quantitative or qualitative. The major gene is assumed to be in Hardy-Weinberg equilibrium following the rules of mendelian transmission through generations. Abel and Bonney (55) introduced survival analysis concepts into the regressive models to take into account variable ages at trait diagnosis. The model of Gauderman and Thomas (54) is an extension to the Cox proportional hazards model and includes measured covariates and unobserved covariates for the genetic component adjusting for variable age at diagnosis for the trait. Two unobserved latent variables are considered: a diallelic major gene and a polygenic component. The major gene is defined as in Bonney's model, and the polygene is assumed to have a standard normal distribution in the population with offspring polygenotypes having expectation equal to the midpoint between their parent's polygenotype with additive variance of half the additive polygenotypic variance. This model, as in Bonney's model, allows for assessing environmental and genetic factors and their interactions. The difference between the two models in terms of explanatory variables is in residual family likeness for the studied trait which is modeled as an additive polygenic component in the model of Gauderman and Thomas and by family dependencies in the original model of Bonney. As such, an extra interaction between the environmental factor(s) and the polygene may be taken into account in the Gauderman and Thomas approach. The other models listed above were proposed for studying quantitative traits (50, 51).

The power to detect gene-environment interaction has been assessed for only the model of Gauderman and Thomas. Power has been succinctly studied using data simulations where the genetic relative risk due to the dominant major gene was 7.4, the environmental relative risk was 2.72, and the allele frequency was 0.1. The authors found 83 percent power to detect an interaction relative risk of 2.7 or more, but only 23 percent power for an interaction risk of 1.6. These simulation studies showed that gene-environment interaction could be estimated well as long as its effect on trait occurrence was large. However, other simulation studies appear necessary to assess the power in detecting gene-environment interaction within a larger spectrum of models by varying mode of inheritance,

allele frequency, and relative risks associated with the environmental factor(s), major gene, and interaction effect.

Segregation analyses have been rarely used to detect gene-environment interaction in human diseases. The major limitation is the need to collect individual data on the trait of interest and environmental covariate data on all family members. Few investigators have been able to collect such data. Gauderman et al. (56) applied their model to a data set of lung cancer families to assess the joint effect of smoking, a major gene, and their interaction on lung cancer risk. The results showed evidence for a mendelian gene segregating in 337 extended pedigrees. A gene-tobacco smoking interaction did not significantly improve the fit of the model indicating that on a multiplicative scale, tobacco and the major gene independently influenced lung cancer risk.

Four groups have used the regressive models to assess gene-environment interactions (57-60). Dizier et al. (58) used the regressive models to better understand the genetic mechanisms controlling immunoglobulin E in allergy in 234 Australian nuclear families. Segregation analysis of immunoglobulin E levels taking into account specific response to allergens was performed. The results showed that familial transmission of immunoglobulin E level was compatible with the segregation of a recessive major gene and residual familial correlation with no significant interaction between the major gene and the specific response to allergens. Andrieu and Demenais (60) assessed gene-reproductive factors interaction in breast cancer risk in 288 families. The results showed evidence for the segregation of a dominant gene with an additional sister-sister dependence as residual likeness between relatives and an interaction between the major gene and parity. A protective effect on breast cancer risk of high parity was observed in nonsusceptible women, but not in susceptible women, indicating that, on a multiplicative scale, parity and the major gene interact to influence breast cancer risk. However, further methodological studies are needed to evaluate the statistical efficiency of the regressive models to detect gene-environment interaction.

#### **Family study design and combined segregation and linkage analyses**

Recently, Gauderman and Faucett (61) proposed adding a linked marker to the studied trait in a joint segregation and linkage analysis to provide additional information on the genetic component of the trait for detecting gene-environment interaction. They considered a continuous trait with a penetrance function based on a linear model that included an intercept and

regression coefficients of covariates (among which was the major gene effect). They assumed that a major gene accounted for all the within-family correlation in the trait. They performed simulations to determine whether adding a linked marker to a segregation analysis improves the efficiency for estimating and power for detecting a gene-environment interaction. They varied interaction strength, recombination fraction of the marker with the studied trait, level of heterozygosity at the marker locus, allele frequency for the major gene, and mode of inheritance and data structures. They showed that including a linked marker in a joint segregation and linkage analysis leads to less bias and increased efficiency for estimating gene-environment interaction effects, and to greater power for detecting interaction compared with segregation analysis alone. However, power gains of the joint analysis never exceeded 9.5 percent compared with the power from the segregation analysis alone. This increase was observed for a closely linked marker ( $\theta = 0.001$ ) with 20 alleles, a high heterozygosity, and a strong interaction. When the interaction was weaker, the gain in power did not exceed 3 percent. Gain in power for the combined segregation and linkage analyses must be considered in conjunction with the extra cost for collecting blood samples and typing the linked markers. As such, this type of study design as a second step in understanding disease etiology is probably not optimal in terms of cost efficiency. However, combined segregation and linkage analyses may be attempted initially, providing a single tool for estimating the effects of trait genes, finding their locations, and assessing gene-environment interaction. Efficiency and power for

such an approach needs to be further assessed according to the transmission pattern of disease, frequency of the disease and exposures, and costs associated with obtaining biologic samples of relatives.

## CONCLUSION

The methods reviewed in this presentation are summarized in table 1. For each method, available risk estimates and required types of subjects are described. Most methods allow for estimating risk associated with a genetic factor, environmental exposure factor, and interaction effect. Case-only studies allow for assessing interaction effects only. While using case-only study designs is easily understood in tumor studies, one may question the utility of this approach in studies involving germline genetic markers (that can be measured in controls). If the exposure risk under study is already well known in the general population, such as smoking in lung cancer, such a study design can be justified, but only if the susceptibility gene involved in the interaction is rare. If the exposure risk is not already well known, which may be the case when the susceptibility gene involved in the interaction is common, the usefulness of case-only study designs may be questionable. Assessment of gene-environment interaction effects without knowing about the gene and environmental main effects would be of little use for public health or individual risk assessment. When the genetic factor is unknown, case-control study designs using related and unrelated controls permit estimation of risks associated with envi-

TABLE 1. Methods where gene-environment interaction can be assessed

Study design	Risk estimate available for:			Information on environmental exposure needed for:			
	Genetic factor	Environmental exposure	Interaction	Cases	Relatives		Unrelated controls
					Affected	Unaffected	
<i>Genetic factor known</i>							
Case-only	No	No	Yes	Yes	No	No	No
Sib-pair, affected-pedigree-member	No	No	No*	Yes	Yes	No	No
Case-control using unrelated control	Yes	Yes	Yes	Yes	No	No	Yes
Two-stage case-control	Yes	Yes	Yes	Yes	No	No	Yes
Case-control using related control	Yes	Yes	Yes	Yes	No	Yes	No
Case-parental study	Yes	No	Yes	Yes	No	Yes	No
Extended sib-pair	Yes	Yes	Yes	Yes	Yes	Yes	No
Combined linkage and segregation analysis	Yes	Yes	Yes	Yes	Yes	Yes	No
<i>Genetic factor unknown</i>							
Case-control using related and unrelated controls	No	Yes	No*	Yes	No	Yes	Yes
Segregation analysis	Yes	Yes	Yes	Yes	Yes	Yes	No
Combined linkage and segregation analysis	Yes	Yes	Yes	Yes	Yes	Yes	No
Twins studies	Yes	Yes	Yes	Yes	Yes	Yes	No

\* May give suggestion for interaction.

ronmental exposures only but may give suggestions for gene-environment interaction.

Efficiency and power for gene-environment interaction assessment have been rarely studied for the various methods presented, and further investigations are needed to define the efficiency spectra of each method in gene-environment interaction assessment. Most reviewed methods seem to be inefficient for detecting interaction of a rare event. Multistage studies could be an alternative approach if the rare event is easily and inexpensively measured. As such, at present, it will likely be too costly to use this approach to study rare genetic factors. There does not appear to be a universal method for assessing gene-environment interaction. The most appropriate approach will depend on the disease, environmental exposure, susceptibility gene, and interaction effect, their values and frequencies, and how much information is available on each one. Further assessment of existing methods and development of new methodologic approaches are needed to improve the ability to detect gene-environment interaction in complex disease.

#### ACKNOWLEDGMENTS

This project was supported by INSERM, the US National Cancer Institute, the Fondation de la Recherche Médicale, and the Association pour la Recherche contre le Cancer. This work was conducted while Nadine Andrieu was a guest researcher at the Genetic Epidemiology Branch of the National Cancer Institute.

The authors thank Drs. Duncan Thomas and Allan Hildesheim for their helpful discussions and suggestions.

#### REFERENCES

- Tiret L, Abel L, Rakotovo R. Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits. *Genet Epidemiol* 1993;10:581-6.
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153-62.
- Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev* 1994;3:173-5.
- 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.
- Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;13:356-65.
- Lehrer S, Sanchez M, Song HK, et al. Oestrogen receptor B-region polymorphism and spontaneous abortion in women with breast cancer. *Lancet* 1990;335:622-4.
- Breslow NE, Day NE. Statistical methods in cancer research. Vol 1—The analysis of case-control studies. Lyon, France: International Agency for Research on Cancer, 1980. (IARC scientific publications no. 32).
- Khoury MJ, Adams MJ Jr, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet* 1988;42:89-95.
- Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med* 1997;16:1731-43.
- Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3-19.
- Amos CI, Elston RC, Wilson AF, et al. A more powerful robust sib-pair test of linkage for quantitative traits. *Genet Epidemiol* 1989;6:435-49.
- Weeks DE, Lange K. The affected-pedigree-member method of linkage analysis. *Am J Hum Genet* 1988;42:315-26.
- Ottman R. Gene-environment interaction: definitions and study designs. *Prev Med* 1996;25:764-70.
- Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med* 1983;2:243-51.
- Thomas DC, Greenland S. The efficiency of matching in case-control studies of risk-factor interactions. *J Chronic Dis* 1985;38:569-74.
- Hwang SJ, Beaty TH, Liang KY, et al. Minimum sample size estimation to detect gene-environment interaction in case-control designs. *Am J Epidemiol* 1994;140:1029-37.
- Goldstein AM, Falk RT, Korczak JF, et al. Detecting gene-environment interactions using a case-control design. *Genet Epidemiol* 1997;14:1085-9.
- Bouchardy C, Benhamou S, Dayer P. The effect of tobacco on lung cancer risk depends on CYP2D6 activity. *Cancer Res* 1996;56:251-3.
- Bain C, Speizer FE, Rosner B, et al. Family history of breast cancer as a risk indicator for the disease. *Am J Epidemiol* 1980;111:301-8.
- Brinton LA, Hoover R, Fraumeni JF Jr. Interaction of familial and hormonal risk factors for breast cancer. *J Natl Cancer Inst* 1982;69:817-22.
- Parazzini F, La Vecchia C, Negri E, et al. Menstrual and reproductive factors and breast cancer in women with family history of the disease. *Int J Cancer* 1992;51:677-81.
- Sellers TA, Potter JD, Severson RK, et al. Difficulty becoming pregnant and family history as interactive risk factors for postmenopausal breast cancer: the Iowa Women's Health Study. *Cancer Causes Control* 1993;4:21-8.
- Colditz GA, Rosner BA, Speizer FE. Risk factors for breast cancer according to family history of breast cancer. *J Natl Cancer Inst* 1996;88:365-71.
- Breslow NE. Case-control study, two-phase. In: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Vol 1. New York, NY: Wiley, 1998:532-40.
- Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect* 1994;102(Suppl 8):47-51.
- Langholz B, Borgan Ø. Counter-matching: a stratified nested case-control sampling method. *Biometrika* 1995;82:69-79.
- Steenland K, Deddens JA. Increased precision using counter-matching in nested case-control studies. *Epidemiology* 1997;8:238-42.
- Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol* 1988;128:1198-206.
- Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988;75:11-20.
- Goldstein AM, Hodge SE, Haile RWC. Selection bias in case-control studies using relatives as the controls. *Int J Epidemiol* 1989;18:985-9.
- Robins J, Pike M. The validity of case-control studies with nonrandom selection of controls. *Epidemiology* 1990;1:273-84.
- Gladden BC. Matched-pair case-control studies when risk factors are correlated within the pairs. *Int J Epidemiol* 1996;25:420-5.

33. Andrieu N, Goldstein AM. Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. *Int J Epidemiol* 1996;25:649-57.
34. Witte JS, Gauderman WJ, Elston RC, et al. Asymptomatic bias and efficiency in case-control studies of candidate genes and gene-environment interactions. I. Basic family designs. *Am J Epidemiol* (in press).
35. Khoury MJ, Flanders WD, Lipton RB, et al. Commentary: the affected sib-pair method in the context of an epidemiologic study design. *Genet Epidemiol* 1991;8:277-82.
36. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet* 1995;57:455-64.
37. Maestri NE, Beaty TH, Hetmanski J, et al. Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. *Am J Med Genet* 1997;73:337-44.
38. Khoury MJ, James LM. Population and familial relative risks of disease associated with environmental factors in the presence of gene-environment interaction. *Am J Epidemiol* 1993;137:1241-50.
39. Andrieu N, Demenais F. Role of genetic and reproductive factors in breast cancer. (Abstract). *Genet Epidemiol* 1994;11:285.
40. Laing AE, Bonney GE, Adams-Campbell L, et al. Reproductive and lifestyle risk factors for breast cancer in African-American women. (Abstract). *Genet Epidemiol* 1994;11:300.
41. Ottman R. Epidemiologic analysis of gene-environment interaction in twins. *Genet Epidemiol* 1994;11:75-86.
42. Ramakrishnan V, Goldberg J, Henderson WG, et al. Elementary methods for the analysis of dichotomous outcomes in unselected samples of twins. *Genet Epidemiol* 1992;9:273-87.
43. Morton NE, MacLean CJ. Analysis of family resemblance. 3. Complex segregation of quantitative traits. *Am J Hum Genet* 1974;26:489-503.
44. Lalouel JM, Morton NE. Complex segregation analysis with pointers. *Hum Hered* 1981;31:312-21.
45. Lalouel JM, Rao DC, Morton NE, et al. A unified model for complex segregation analysis. *Am J Hum Genet* 1983;35:816-26.
46. Sellers TA, Bailey-Wilson JE, Elston RC, et al. Evidence for Mendelian inheritance in the pathogenesis of lung cancer. *J Natl Cancer Inst* 1990;82:1272-9.
47. Gilligan SB, Borecki IB. Examination of heterogeneity in 200 Danish breast cancer pedigrees. *Genet Epidemiol Suppl* 1986;1:67-72.
48. Morton NE. The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am J Hum Genet* 1956;8:80-96.
49. Hodge SE, Anderson CE, Neiswanger K, et al. The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. *Am J Hum Genet* 1983;35:1139-55.
50. Moll PP, Sing CF, Lussier-Cacan S, et al. An application of a model for a genotype-dependent relationship between a concomitant (age) and a quantitative trait (LDL cholesterol) in pedigree data. *Genet Epidemiol* 1984;1:301-14.
51. Konigsberg LW, Blangero J, Kammerer CM, et al. Mixed model segregation analysis of LDL-C concentration with genotype-covariate interaction. *Genet Epidemiol* 1991;8:69-80.
52. Bonney GE. Regression logistic models for familial disease and other binary traits. *Biometrics* 1986;42:611-25.
53. Demenais F, Lathrop M. REGRESS: a computer program including the regressive approach into the LINKAGE programs. (Abstract). *Genet Epidemiol* 1994;11:291.
54. Gauderman WJ, Thomas DC. Censored survival models for genetic epidemiology: a Gibbs sampling approach. *Genet Epidemiol* 1994;11:171-88.
55. Abel L, Bonney GE. A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. *Genet Epidemiol* 1990;7:391-407.
56. Gauderman WJ, Morrison JL, Carpenter CL, et al. Analysis of gene-smoking interaction in lung cancer. *Genet Epidemiol* 1997;14:199-214.
57. Thein SL, Sampietro M, Rohde K, et al. Detection of a major gene for heterocellular hereditary persistence of fetal hemoglobin after accounting for genetic modifiers. *Am J Hum Genet* 1994;54:214-28.
58. Dizier MH, Hill M, James A, et al. Detection of a recessive major gene for high IgE levels acting independently of specific response to allergens. *Genet Epidemiol* 1995;12:93-105.
59. Briollais L, Chompret A, Guilloud-Bataille M, et al. Genetic and epidemiological risk factors for a malignant melanoma-predisposing phenotype: the great number of nevi. *Genet Epidemiol* 1996;13:385-402.
60. Andrieu N, Demenais F. Interactions between genetic and reproductive factors in breast cancer risk in a French family sample. *Am J Hum Genet* 1997;61:678-90.
61. Gauderman WJ, Faucett CL. Detection of gene-environment interactions in joint segregation and linkage analysis. *Am J Hum Genet* 1997;61:1189-99.