Review

Epidemiological methods for studying genes and environmental factors in complex diseases

David Clayton, Paul M McKeigue

Exploration of the human genome presents new challenges and opportunities for epidemiological research. Although the case-control design is quicker and cheaper for study of associations between genotype and risk of disease than the cohort design, cohort studies have been recommended because they can be used to study gene-environment interactions. Although the scientific relevance of statistical interaction is pertinent, the main disadvantage of the case-control design—susceptibility to bias when estimating effects of exposures that are measured retrospectively—does not necessarily apply when studying statistical interaction between genotype and environmental exposure. Because correctly designed genetic association studies are equivalent to randomised comparisons between genotypes, conclusions about cause can be drawn from genetic associations even when the risk ratio is modest. For adequate statistical power to detect such modest risk ratios, the case-control design is more feasible than the cohort design.

A key objective of research in human genetics is to advance knowledge of how genetic and environmental factors combine to cause disease. New opportunities for epidemiology have been heralded: for example, Shpilberg and colleagues[1] stated that "The sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street pump". To take advantage of these opportunities, several countries are establishing collections of DNA samples with data on clinical outcome. In the UK, a cohort study of 500 000 individuals over 10 years is planned. In this report, we review the basis for the new optimism, and examine the methodological issues that arise in the design of epidemiological studies based on DNA collections, with emphasis on the choice between cohort and case-control designs.

One of the main contributions of epidemiology to research methods has been the development of the case-control design for study of effects of exposures—defined broadly to include behaviours and physiological measurements—on risk. This study design is based on the principle that any risk ratio that can be estimated in a cohort study can also be estimated in a case-control study. With a case-control study, more precise characterisation of outcomes is possible than would be feasible with a cohort design; for instance, in a case-control study of stroke, investigators can ensure that all cases are scanned to distinguish haemorrhagic from ischaemic strokes. With a given outlay of resources, far greater statistical power to detect associations can be achieved with a case-control design than with a cohort design. As we shall see later, this is of great importance in genetic epidemiology.

Lancet 2001; **358:** 1356–60

Department of Medical Genetics, Cambridge University, Level 4, Cambridge Institute for Medical Research, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK (D Clayton MA)**; and Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London** (Prof P M McKeigue PhD)

Correspondence to: Mr David Clayton
(e-mail: david.clayton@cimr.cam.ac.uk)

In both case-control and cohort studies, measurement of exposures such as diet is subject to error. The main disadvantage of case-control studies, compared with cohort studies, is that these measurement errors can differ systematically between cases and controls. This difference can give misleading results. For example, if people diagnosed with cancer recall their past dietary fat intake differently from controls, estimates of the association between dietary fat and cancer will be biased.[2] Because this problem does not arise in measurement of genotype, the case-control design is the method of choice when the objective is simply to study associations between genotype and disease risk.

The rationale for setting up cohort studies of genetic effects on disease risk is based on the argument that, because cohort studies can measure environmental exposures before disease onset, they are better than the case-control design for study of gene-environment interactions. Study of such interactions is thought to make detection of genes that influence disease risk easier, to allow individuals at high risk to be identified for targeted intervention, and to advance understanding of biological pathways leading to disease.

To question the emphasis on gene-environment interaction might seem perverse, since disease clearly arises from the interplay of these factors. However, advocates of this approach seldom define what they mean by "interaction". Despite current enthusiasm for study of gene-environment interactions, the closely related issue of how to define and interpret interaction between environmental factors remains unresolved after two decades of debate. Similar difficulties arise in the study of interaction between genes (epistasis).[3] The scientific value of focusing on gene-environment interactions has not been established, and in any case, the technical advantages of cohort studies over case-control studies in detection of statistical interactions between genetic and environmental effects are less clear than has been assumed. We suggest that epidemiologists should focus instead on use of genetic associations to test hypotheses about causal pathways amenable to intervention.

For personal use. Only reproduce with permission from The Lancet Publishing Group.

Statistical and biological interaction

The meaning of the term "interaction" can be a cause of confusion. Because of this ambiguity, statisticians commonly preface their discussion of interaction with a disclaimer that statistical interaction should not be confused with biological or causal interaction.[4] Cox[5] noted that "The notion of interaction and indeed the very word itself are widely used in scientific discussion. This is largely due to the relation between interaction and causal connexion. Interaction in the statistical sense has, however, a more specialized meaning related, although often in only a rather vague way, to the more general notion."

In statistical terms, gene-environment interaction is present when the effect of genotype on disease risk depends on the level of exposure to an environmental factor, or vice versa. This definition depends on how effects on risk are measured. The most usual measure of effect in epidemiology is the ratio of disease incidence between exposed and unexposed individuals, which, in correctly designed case-control studies, can be measured by an odds ratio. With this definition, no interaction corresponds with a multiplicative model for the joint effects of two or more risk factors, in which the risk ratio between individuals exposed and unexposed to risk factor A does not vary over strata defined by exposure to another risk factor B. Statistical interaction is defined as lack of fit to this multiplicative model.

If we were to define the measure of effect as a rate difference, interaction would be defined as lack of fit to an additive model for the joint effects of the two risk factors. For example, in women, the risk of venous thrombosis is increased about eight-fold in those with the Arg506Gly (Leiden) mutation in the factor V gene, and four-fold in oral contraceptive users, compared with women who have neither risk factor.[6] Table 1 shows how the joint effects would differ according to whether the model is additive or multiplicative. With a multiplicative model, lack of interaction would imply a 32-fold increase of risk to women with both risk factors, compared with women with neither risk factor. With an additive model, lack of interaction would imply an 11-fold increase in risk to women with both risk factors. As it happens, the observed risk ratios approximate to a multiplicative model.[6]

The general issue addressed by the idea of statistical interaction is the quantitative description of joint effects. Simple working models are usually chosen on empirical grounds, and the adequacy of these models is assessed by tests for interaction. When testing for the average effect of a risk factor (main effect), the null hypothesis of no difference between risk in exposed and unexposed individuals has clear biological interpretation. When testing for interaction, the null hypothesis is that the joint action of two factors on incidence is described by a mathematical model. If this null hypothesis has no obvious biological interpretation, testing for statistical interaction might not contribute to biological understanding.

Oral contraceptive use	Leiden mutation			
	Multiplicative model		Additive model	
	No	Yes	No	Yes
No	1	8	1	8
Yes	4	32	4	11

Data show risk per 10 000 woman-years.

Table 1: **Examples of multiplicative and additive models for effects of genotype and oral contraceptive use on risk of venous thrombosis (hypothetical data)**

The relation between biological models of mechanism and statistical models for the joint effects of risk factors has been explored in detail by epidemiologists, but investigators eventually recognised that the same statistical model for disease risk could be obtained from many different models of mechanism. Thompson[4] concluded that, "Unfortunately, choice among theories of pathogenesis is enhanced hardly at all by epidemiological assessment of interaction . . . What few causal systems can be rejected on the basis of observed results would provide decidedly limited etiological insight."

Estimation of statistical interactions as a basis for targeting interventions

One argument for trying to obtain quantitative estimates of the joint effects of genotype and environment is that these estimates provide a basis for targeting interventions at individuals at high risk.[7] However, the rationale for targeted intervention does not generally depend on the ability to detect statistical interaction in terms of lack of fit to a multiplicative model. In a multiplicative model, lack of statistical interaction between genotype and an environmental risk factor implies that the benefit of avoiding exposure to the environmental risk factor will be greater for individuals with a high-risk genotype than for those with a low-risk genotype. Thus, with the example of venous thrombosis and oral contraceptive use, a multiplicative model implies that the excess risk to oral contraceptive users compared with non-users will be eight times greater in women with the Leiden mutation than in those without this mutation. Unless there was compelling evidence to reject a multiplicative model in favour of an additive model, the demonstration of an association between venous thrombosis and the Leiden mutation would lead us to conclude that women with this high-risk genotype should avoid known environmental risk factors such as oral contraceptives.

Outside the realm of therapeutics, the practical usefulness of environmental interventions targeted at people in accordance with their genotype is likely to be limited. Rose[8] argued that (at least for non-communicable disease control) the predicted health gains are generally greater from interventions that are directed at the whole population than from those targeted at a high-risk group. This phenomenon is due to most cases of multifactorial disease arising in individuals who do not fall into a targeted high-risk group, and because individuals who are at high risk find adoption of behaviours that deviate from population norms difficult.

Effect of considering gene and environment on statistical power

If disease is caused by an interplay of genetic and environmental factors, then, plausibly, studies will be more powerful if they measure both types of factor and model their joint effects in the analysis; this assumption, however, is too simplistic. Even if there are subgroups of genetically susceptible individuals, and the size of effect associated with an environmental exposure varies with genotype, the direction of this effect is unlikely to vary with genotype. This lack of variation limits the gain in statistical power obtained by fitting a model that allows the effect of the environmental factor on disease risk to vary with genotype, rather than simply testing for overall association between the environmental exposure and disease. If we could specify in advance that the effect of the environmental factor on disease risk would be restricted to a subgroup of individuals with a particular genotype, there would, of course, be a gain in power from

For personal use. Only reproduce with permission from The Lancet Publishing Group.

testing only this subgroup for the effect of the environmental factor. In practice, such an extreme situation is unlikely to be frequently encountered in the study of complex diseases, and entails a level of knowledge of underlying biology which would probably render epidemiological studies redundant. In less extreme situations, and where previous knowledge is more limited, a combined test would need to be done for the main effect of environmental exposure and its interaction with genotype. Since such tests have multiple degrees of freedom, the gain in power is much reduced; indeed, power might even be lost.

Usefulness of cohort studies in the study of statistical interactions

We have argued that study of statistical interactions between genetic and environmental factors in epidemiological studies is, perhaps, not as interesting as it might seem at first sight. Nevertheless, the quantification of joint effects remains a legitimate aim of such studies. Detection of departure from the widely used model of multiplicative effects provides some interesting lessons. In particular, we can re-examine the presumption that cohort studies are better than case-control studies for this purpose.

To develop the argument, imagine that we can do a perfect case-control study of an environmental factor, and that we can further classify individuals as genetically susceptible or not. The hypothetical results of such a study are shown in table 2, in which the letters *a* to *h* represent cell frequencies. From each 2×2 table we can calculate the ratio of the odds of disease in exposed individuals to the odds of disease in unexposed individuals. Thus, the rate ratio for exposure to the environmental factor is estimated by the odds ratio *ad*/*bc* in the genetically susceptible group, and by the odds ratio *eh*/*fg* in the other group. Interaction, with respect to the multiplicative model, contrasts these two rate ratios, and is measured by their ratio, *adfg*/*bceh*. As noted earlier, if environmental exposure is subject to different measurement errors in cases and controls, then the odds ratios *ad*/*bc* and *eh*/*fg*—estimating the effect of the environmental factor—will be distorted. We could postulate that the ratio would likewise be distorted, so that any conclusion about gene-environment interaction would be unsafe.

Consider the same data, rearranged as in table 3. We may now calculate the odds ratio *ag*/*ce* to measure association between genotype and environmental exposure in cases, and the odds ratio *bh*/*df* to measure association between genotype and environmental exposure in controls. The interaction parameter in the table is the ratio *agdf*/*cebh* of these two odds ratios—exactly the same as before. This rearrangement casts a new slant on the problem of estimation of statistical interaction. If genotype and environmental exposure are assumed to be independent in the population, and the disease is rare, such that disease-free controls are assumed to be representative of the population, then we can expect the second odds ratio *bh*/*df* to be 1. Interaction is then estimated simply from the association between genotype and environmental exposure in cases, measured by the odds ratio *ag*/*ce*. This procedure is the basis of the "case-only" design, which has been proposed to study gene-environment interaction, defined as deviation from a multiplicative model for the joint effects of genotype and environmental exposure.[9] The assumption that genotype and environmental exposure are independent in the population under study can easily be tested in a control group.

Although the usefulness of the case-only design remains to be established, the argument on which the design is based has other lessons. In particular, it clarifies the effect of errors in measurement of environmental exposure. Since the interaction effect can be estimated from cases only, different measurement error between cases and controls is not a serious problem. Although there would be a problem if errors in measurement of environmental exposure differed with genotype, generally this problem is unlikely to occur, except in behavioural genetics in which genotype might influence how people report their exposure. Measurement errors that are independent of genotype will simply bring the odds ratios in table 2 closer to 1. However, with the assumption that genotype and environment are independent in the population, the second odds ratio is 1 already, and it follows that the only effect of measurement error is to bring the interaction parameter closer to 1—ie, towards the hypothesis of multiplicative action of gene and environment. Even if measurement error is greater in a case-control design (in which environmental exposure is usually measured retrospectively) than in a cohort design (in which measurements can be made at the time of exposure), the case-control study will generally have greater statistical power to detect gene-environment interaction because a much larger number of cases can be studied for a given outlay of resources.

Even when quantification of the joint effects of gene and environment is important, if errors in measurement of either factor have been made, the form of the relation will be distorted, being biased towards a model of multiplicative action. If interaction is defined as lack of fit to a multiplicative model, a test for interaction will be conservative, such that if the null hypothesis is correct, the test will not yield significant results more often than those expected by chance. With any other definition of interaction, tests for interaction will not necessarily be conservative in the presence of measurement error. This reason might explain why multiplicative models are usually an adequate fit to the observed data in practice. In any case, it casts further doubt on the scientific interest of the study of statistical interactions.

Using genetic associations to test hypotheses about causal pathways

Optimism about prospects for epidemiology in the post-genome era contrasts with a pessimistic and widely quoted view that modern epidemiology "faces its limits".[10] Most epidemiological research now focuses on attempts to

Environmental exposure	Positive genotype		Negative genotype	
	Cases	Controls	Cases	Controls
Yes	*a*	*b*	*e*	*f*
No	*c*	*d*	*g*	*h*
Odds ratio	$\frac{ad}{bc}$		$\frac{eh}{fg}$	

Table 2: **Odds ratios for association of disease with environmental exposure, by genotype**

Environmental exposure	Cases		Controls	
	Positive genotype	Negative genotype	Positive genotype	Negative genotype
Yes	*a*	*e*	*b*	*f*
No	*c*	*g*	*d*	*h*
Odds ratio	$\frac{ag}{ce}$		$\frac{bh}{df}$	

Table 3: **Odds ratios for association between environmental exposure and genotype, in cases and controls separately**

For personal use. Only reproduce with permission from The Lancet Publishing Group.

estimate modest risk ratios associated with environmental or behavioural exposures that cannot be measured accurately. In this situation, standard techniques for control of bias and confounding become untrustworthy because the effects under study are small in relation to the unavoidable biases of epidemiological studies. Thus, observational epidemiology is increasingly unable to resolve questions of major public health importance, such as the relation of diet to cancer risk.

One of the most important contributions of genetic epidemiology could be the ability to overcome limitations of classic epidemiological techniques, through "Mendelian randomisation".[11] In a correctly designed genetic association study, the laws of Mendelian genetics ensure that comparison of groups of individuals defined by genotype is equivalent to a randomised comparison, since these groups will not differ systematically, except with respect to allelic associations (linkage disequilibrium) that extend over a short genomic region from the locus under study. Mendelian randomisation is most easily appreciated in study designs that test for dependence of outcome (in offspring) on alleles transmitted from parents who are heterozygous at the locus under study, as in the transmission-disequilibrium test.[12] This test is equivalent to a randomised trial, such that each of the two alleles in a parent has an equal chance of being transmitted to the offspring.

This argument can be extended to ordinary case-control designs, in which parents are not genotyped; controlling for population substructure (stratification of the population into subpopulations that have different allele frequencies) in such studies is sufficient to eliminate confounding by alleles at unlinked loci or environmental factors. Unknown population substructure can be estimated and controlled with genotype data from a panel of markers unlinked to the locus under study.[13,14] By contrast with epidemiological studies of behavioural risk factors, for which bias and residual confounding are difficult to exclude, an association between genotype and outcome in a correctly designed study cannot be attributable to bias or residual confounding (except by alleles at nearby loci). The main problem is to exclude chance as an explanation, which can be achieved simply with a larger sample size. With stringent thresholds for declaration of significance and control for population substructure, even a modest risk ratio in a genetic association study is compelling evidence for a causal relation.

One application of this approach is to study the effect of genetic polymorphisms that affect the pathway of interest. For example, concentrations of fibrinogen in plasma have consistently been found to predict coronary disease,[15] but investigators have not established whether this association has a causal basis. This association has been investigated by study of a polymorphism in the β-fibrinogen gene that is known to influence concentrations of fibrinogen in plasma.[11] Comparison of coronary disease risk in individuals with none, one, or two copies of the allele that results in increased fibrinogen concentrations can be interpreted as an experiment of nature in which individuals have been randomly allocated to high or low fibrinogen concentrations. The effect of genotype on risk can be estimated in a large case-control study that compares the observed risk ratio with the expected risk ratio on the basis of the known relation between coronary risk and plasma fibrinogen concentrations.[11]

If the pathway of interest is the effect of a dietary component, identification of a functional polymorphism that alters the metabolism or bioavailability of this component could be possible. For instance, homozygosity for the (C677T) variant of the methylene tetrahydrofolate reductase (*MTHFR*) gene is associated with reduced folate-dependent enzyme activity that can be partly reversed by dietary folate supplementation.[16] A case-control study of the relation between the TT genotype and risk of neural tube defect[17] can be interpreted as equivalent to a randomised trial of the effect on disease risk of alteration of the availability of folate.

These examples demonstrate one use of genetic association studies. To test whether a risk factor has a causal relation to disease risk, we can look for polymorphisms that affect the risk factor or the metabolic pathway on which the action of the risk factor depends, then examine the effects of these polymorphisms on disease risk in a large case-control study. The ability of Mendelian randomisation to eliminate bias and residual confounding allows us to examine the effects associated with genetic polymorphisms, even when these effects are small. Of special interest are polymorphisms that alter the metabolism of a dietary substrate or the activity of an enzyme or receptor that is a potential drug target. Study of polymorphisms that disturb pathways that are not readily amenable to intervention (such as HLA antigens) is less interesting, even when these polymorphisms are associated with large risk ratios.

More generally, discovery of new genetic associations could tell us which exposures to look for in the environment. For instance, some studies suggest that polymorphisms in the *N*-acetyltransferase gene, that affect the activity of the *N*-acetylation pathway, could influence the risk of colon cancer.[18] This association, if confirmed, provides a basis for investigation of dietary constituents that are substrates for the acetylator pathway, such as the heterocyclic amines in cooked meat. In this example, as with the *MTHFR* gene, there is a possible biological interaction between genotype and dietary intake, but testing for statistical interactions between genotype and dietary intake would not contribute much to our understanding of these biological interactions or to our ability to exploit them in disease prevention.

A crucial requirement of this approach is that the study design should have adequate statistical power to confirm or exclude modest risk ratios, because polymorphisms that have large effects on risk factors are rarely available. For example, when studying the effect of polymorphisms in the β-fibrinogen gene, the effect of the allele that increases plasma fibrinogen concentrations is equivalent to only 20% of the within-group SD of fibrinogen concentration, and a sample of more than 4000 cases was required to confirm or exclude the predicted risk ratio for coronary disease of 1·2.[11] This number of cases is far larger than the number required to detect the relation of plasma fibrinogen to risk of coronary disease in a prospective study.[15] The requirement that modest risk ratios should be detected, and stringent criteria for statistical significance should be adopted when large numbers of loci are tested, necessitates studies more powerful than any hitherto considered. Since cohort studies sufficiently large for this purpose are unlikely to be practicable, except for a few common diseases, proposals for very large cohort studies of genetic associations should be critically examined against alternatives.

The prospects for epidemiology in the post-genome era depend on understanding how to use genetic associations to test hypotheses about causal pathways, rather than on modelling the joint effects of genotype and environment.

For personal use. Only reproduce with permission from The Lancet Publishing Group.

D Clayton is supported by a Wellcome Trust/Juvenile Diabetes Research Foundation Principal Research Fellowship. P McKeigue is supported in part by NIH/NIMH grant MH60343.

References

1 Shpilberg O, Dorman JS, Ferrell RE, Trucco M, Shahar A, Kuller LH. The next stage: molecular epidemiology. *J Clin Epidemiol* 1997; **50:** 633–38.

2 Giovannucci E, Stampfer MJ, Colditz GA, et al. A comparison of prospective and retrospective assessments of diet in the study of breast cancer. *Am J Epidemiol* 1993; **137:** 502–11.

3 Cordell HA, Todd JA, Hill NJ, et al. Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 2001; **158:** 357–67.

4 Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 1991; **44:** 221–32.

5 Cox DR. Interaction. *Int Stat Rev* 1984; **52:** 1–31.

6 Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet* 1994; **344:** 1453–57.

7 Khoury MJ, Wagener DK. Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors. *Am J Hum Genet* 1995; **56:** 835–44.

8 Rose G. The strategy of preventive medicine. Oxford: Oxford University Press, 1992.

9 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.

10 Taube H. Epidemiology faces its limits. *Science* 1995; **269:** 163–81.

11 Youngman LD, Keavney BD, Palmer A, et al. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: test of causality by "Mendelian randomisation". *Circulation* 2000; **102** (suppl II)**:** 31–32.

12 Julier C, Hyer RN, Davies J, et al. Insulin–IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility. *Nature* 1991; **354:** 155–59.

13 McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations by conditioning on parental admixture. *Am J Hum Genet* 1998; **63:** 241–51.

14 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000; **67:** 170–81.

15 Meade TW, Mellows S, Brozovic M, et al. Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study. *Lancet* 1986; **2:** 533–37.

16 Bailey LB, Gregory JF. Polymorphisms of methylenetetrahydrofolate reductase and other enzymes: metabolic significance, risks and impact on folate requirement. *J Nutr* 1999; **129:** 919–22.

17 Shields DC, Kirke PN, Mills JL, et al. The "thermolabile" variant of methylenetetrahydrofolate reductase and neural tube defects: an evaluation of genetic risk and the relative importance of the genotypes of the embryo and the mother. *Am J Hum Genet* 1999; **64:** 1045–55.

18 Roberts-Thomson IC, Ryan P, Khoo KK, Hart WJ, McMichael AJ, Butler RN. Diet, acetylator phenotype, and risk of colorectal neoplasia. *Lancet* 1996; **347:** 1372–74.

For personal use. Only reproduce with permission from The Lancet Publishing Group.