Study Design in Genetic Epidemiology: Theoretical and Practical Considerations

Alice S. Whittemore, Lorene M. Nelson

Recent advances in molecular genetics have created new opportunities and challenges for genetic epidemiologists. Here we review some of the issues that arise when designing a study involving the genetic epidemiology of chronic diseases of late onset, such as cancer. We discuss two considerations that influence the choice of design. The first consideration is the study’s goals. We describe the goals of identifying new susceptibility genes for a disease, of estimating important characteristics of known genes, and of learning how to prevent the disease in the genetically susceptible. We indicate how these goals affect the choice of design and present some guidelines for choosing designs that effectively address them. The second consideration is the set of practical constraints to successfully conducting the research. These constraints include problems of potential selection bias, reduced response rates, problems particular to family registries, problems particular to the cultures of various ethnic groups, and ethical issues. We indicate how these constraints affect the choice of design and discuss ways to deal with them. [Monogr Natl Cancer Inst 1999;26:61–9]

Technologic developments in molecular genetics have created new opportunities and challenges for genetic epidemiologists. Within the past three decades, it has become feasible to localize and sequence a disease-susceptibility gene before its function is known. Such gene identification is accomplished using one or more of several strategies, collectively known as positional cloning. These strategies include linkage and association studies and, for site-specific cancers, studies of allelic loss and differential gene expression in cancer cells compared with normal cells in the same individual. Positional cloning of genes using one or more of these techniques leads immediately to basic biologic questions. For example, we need to know the function of the protein coded by the gene and how polymorphisms of the gene lead to functional variants in its protein that alter disease risk. Answers to these questions, if available, can suggest important gene–environment or gene–gene interactions and can motivate preventive strategies.

An alternative approach to positional cloning for gene identification is evaluation of candidate genes, i.e., those coding for proteins whose functions are thought to be involved in the disease process. By definition, the function of a candidate gene is at least partially understood. Successful gene identification, whether by positional cloning or by implication of candidate genes, introduces several new areas of inquiry for the genetic epidemiologist. For example, a question of practical clinical importance concerns the age-specific disease risks associated with the various genotypes. A question with less immediate clinical consequences, but one that is important for public health planning and resource allocation, concerns the population frequencies of genotypes associated with increased risk. A related issue is estimation of that fraction of the disease burden due to variant genotypes.

When several genes have been identified, we need to know how these genes interact to affect disease risk. Perhaps the most important issue, from a public health perspective, concerns how lifestyle characteristics modify risk in carriers of high-risk genotypes. Specifically, what can such carriers do to reduce their risk? Why do some carriers remain disease free well into old age? Studies aimed at investigating these areas of inquiry, which we shall call gene characterization studies, present challenging issues for genetic epidemiologists and will continue to do so well into the 21st century.

This paper considers some of the issues involved in designing studies to identify and characterize disease genes. We focus on two considerations: 1) the goals and objectives of the study and 2) the practical constraints to conducting the research. We indicate how these two considerations affect the choice of design, and we summarize some guidelines for choosing designs that effectively address the study objectives while circumventing the limiting constraints.

Effect of Scientific Goals on Design

The goals of a study typically are to identify new genes that predispose to the disease of interest or, if a predisposing gene is known, to estimate the risks associated with specific alleles and the prevalences of those risky alleles and to determine the way that other genes or lifestyle characteristics modify those risks. An additional important goal is to evaluate disease-prevention strategies in the genetically susceptible. Currently, for example, uncertainty exists over the advisability of prophylactic mastectomy or oophorectomy or both in female carriers of the genes BRCA1 and BRCA2, largely because of the lack of data comparing the benefits of such surgery (cancer risk reduction) with its major physical and psychologic costs. However, a retrospective cohort study of 639 women with a family history of breast cancer who underwent bilateral prophylactic mastectomy at the Mayo Clinic between 1960 and 1993 and who were followed for subsequent breast cancer incidence showed a 90% reduction of incidence compared with that predicted in the absence of surgery (1).

Identifying New Genes

Studies aimed at gene identification fall into several categories. These categories include (a) genome-wide scans using
parametric and nonparametric linkage analyses, (b) association studies using cases and unrelated controls, (c) genome-wide scans using family-based association studies, (d) evaluation of candidate genes, and, for cancer, (e) allelic loss and differential gene expression studies based on comparison of cancer cells and normal cells in the same individual. Choosing among these designs requires that we consider several factors, including the anticipated frequencies and penetrances of alleles at the disease locus, the numbers of alleles and frequencies of alleles at nearby markers, and the disease phenotype of interest (e.g., multiple cancers or intermediate biomarkers). Genome-wide linkage and association scans exploit the recent availability of a large set of DNA variants, called markers, each having a known location on a particular chromosome. A marker might be a gene, a sequence of bases that varies from one chromosome to another in a detectable way, or a single nucleotide polymorphism (SNP). The chromosomes of family members with and without the disease are labeled with respect to their observed alleles at each of the markers.

**Genome-Wide Scans Based on Linkage**

Linkage or allele-sharing studies exploit the fact that markers having some alleles shared by the diseased family members and other alleles shared by the nondiseased members are likely to be proximal to a disease-susceptibility gene. When such disease-specific sharing of the marker alleles is unlikely to be a result of chance, the chromosomal region around the marker(s) is searched intensively in an attempt to identify a critical subregion containing the disease gene and, indeed, the gene itself. Association studies, by contrast, search for the correlation between disease and specific marker alleles in much the same way as classic epidemiologic studies search for the correlation between disease and personal attributes.

Parametric linkage analysis requires a genetic model for both the disease phenotype and the marker genotypes. This analysis includes specification of the locations of the markers, with an unknown parameter representing the location of the putative disease gene. This parameter, which is of primary interest, is estimated from data on observed disease phenotypes and marker genotypes in families, by the method of maximum likelihood. In contrast, nonparametric or “model-free” linkage analysis, also called allele-sharing analysis, uses the family data on disease phenotypes and marker genotypes to evaluate departures from the null hypothesis that the marker genotypes are not near a disease gene. Whether parametric or nonparametric, linkage analysis is concerned not with association between disease phenotype and a particular marker allele, but rather with patterns of marker allele sharing within diseased and nondiseased family members. Suppose, for example, families are typed at a marker with alleles A and B that is close to a disease gene. Then allele A may segregate with disease in some of the families, while allele B does so in others; the clues come from patterns of allele segregation with disease rather than association of specific alleles with disease. A more detailed discussion of linkage studies and issues in their design can be found in the paper by Schaid et al. (2) in this monograph.

**Association Studies With Case Patients and Unrelated Control Subjects**

An alternative to linkage for gene identification is a set of statistical methods that we shall call association analyses. The simplest is the classic case–control design that compares marker genotypes of diseased case patients with those of unrelated disease-free control subjects. With current technology, the case–unrelated control design is used to examine a small number of candidate gene polymorphisms as they relate to risk of the disease under study. This type of study design is appropriate for the study of relatively common alleles that confer a modest or small effect on disease risk. In the near future, technology will make it possible to conduct whole-genome association studies that compare allele frequencies across the genome between case patients and unrelated control subjects. Once the catalog of human DNA sequence variation is complete, it will be possible to directly test the common functional variants that occur in coding regions (expecting that two or three variants each with a frequency of 10% or greater will occur in a given coding region) and to compare these allele frequencies between case patients and unrelated control subjects (3). A second approach will involve the use of SNPs in a dense map that spans both the coding and noncoding regions of the human genome (3). This approach will be based on the rationale that affected individuals will be characterized by a particular haplotype in the neighborhood of the susceptibility gene and that comparison with unaffected control subjects will identify specific regions where there is an increased prevalence of a particular set of SNP alleles in affected individuals. The availability of approximately 100 000–200 000 SNPs will allow a dense map for the identification of susceptibility genes; however, the large number of statistical comparisons will require large sample sizes to account for the increased probability of identifying associations that occurred on the basis of chance alone.

Investigators have worried that measures of association obtained from studies that compare case patients and unrelated control subjects may be biased by confounding because of ethnic stratification of the population. Suppose, for instance, that the locus under test is not near a disease locus, but that the population consists of two ethnic groups, with the first group having higher prevalence of both the disease and a specific allele than the second group. A random sample of cases would contain a higher fraction of the first ethnic group than would either the general population or a sample of controls and, thus, a higher total count of the allele. In this situation, the test statistic will lead to rejection of the null hypothesis more often than it should. Thus, failure to account for such ethnic stratification, either by a matched design or by a stratified analysis, could lead to invalid conclusions. Adjusting for population stratification requires collecting detailed race or ancestry information; however, this approach may leave residual confounding because many genetic traits vary in frequency within apparently homogenous ethnic groups. Race or ethnicity is also an incomplete surrogate for genetic makeup because there may be cultural and geographic influences on risk within subgroups of a given racial or ethnic group. Even within an apparently homogenous ethnic group, there may be different allele frequencies at many loci because of the effects of different geographic locations or migration patterns, and both of these factors may affect ethnic admixture within the apparently homogenous group. Thus, it is advisable to compare cases and controls within narrowly defined ethnic groups. Such comparisons may be difficult to achieve in the United States because of the increasing proportion of individuals with mixed race or ancestry. It is worth noting that the potential bias of concern here is a classic example of confounding (by
ences to be due to such bias. These issues are discussed in more detail by Caporaso et al. (5) in this monograph.

**Genome Scans Using Family-Based Association Analyses**

The problem of confounding by ethnic stratification in case-control studies that use unrelated control subjects has prompted interest in family-based designs, such as comparison of genotypes of affected and unaffected siblings or comparison of genotypes of affected offspring with those of their parents (6). These designs automatically match “controls” to cases on ethnic ancestry. For example, much has been written about the transmission disequilibrium test that compares the marker alleles transmitted from heterozygous parents to affected offspring with those not transmitted [see Schaid (7) for a review and commentary]. In this monograph, Gauderman et al. (6) note that family-based designs are generally less efficient than designs based on unrelated control subjects. Research is needed to investigate the trade-offs in bias and efficiency involved in the various design choices for association studies.

An important question in searching for new genes concerns the relative merits of linkage and association analyses for genome scans. Risch and Merikangas (8) argue that, whereas the former have more power for genes with rare, highly penetrant disease-susceptibility alleles such as are often observed among families with autosomal dominant inheritance, association studies are preferred for genes having common alleles with low penetrance. Their conclusions are based on the assumption that the disease-causing and marker polymorphisms are identical. This assumption has been questioned, and consequences of its violation have been investigated (9–11). In situations in which the disease and the marker polymorphisms are not identical even within the same gene, if the extent of disequilibrium between them is not perfect and if the frequencies of alleles at the two loci differ substantially, association analyses can be less powerful than linkage analyses even for common, low-penetration polymorphisms. Thus, the extent of disequilibrium between disease-causing and marker loci is a critical determinant of the power of association studies when genotypes at the disease locus have moderate to low penetrance. Another issue that must be considered when contemplating a genome-wide association scan is the likely etiologic basis of the disease of interest. Complex diseases, in which environmental and behavioral factors combine with genetic factors to cause increased risk, must be investigated in more complex ways than monogenic or oligogenic diseases.

**Evaluation of Candidate Genes**

To avoid inflated type-I error probabilities, the genome scans described above must accommodate the multiple testing involved in evaluating the 300 to 100,000 or more markers evaluated in a scan. In contrast, investigations of candidate genes, based on an a priori biologic hypothesis that a particular gene or family of genes is involved in the disease, avoid this problem. Examples of candidate genes include those involved in estrogen metabolism in relation to breast cancer; in the synthesis, metabolism, and response to testosterone in relation to prostate cancer; and in the metabolism of heterocyclic amines from well-cooked meat in relation to colon cancer, as described by Le Marchand (12) in this monograph. Candidate genes can be examined with the use of the analytic methods described above for association studies. This approach suffers from a drawback similar to that noted above for association-based scans, namely, lack of knowledge of the exact polymorphism that is causally related to disease risk. Instead, investigators identify polymorphisms having variation in the general population and then examine them for association with disease. The absence of tight linkage disequilibrium between causal and measured alleles can induce problems similar to those of misclassification error, namely, power loss and attenuation bias in relative risk estimates.

Some issues in gene identification apply to both genome scans and evaluation of candidate genes. For example, there is a need for designs and analytic methods that accommodate the late and variable ages at onset of most chronic diseases. Clearly, a man who dies in his 40s from a myocardial infarction should not be considered unaffected with respect to diseases of late onset, such as prostate cancer. Analyses that account for age, however crudely, should be used when investigating the etiologic basis of diseases of late onset. A related issue, discussed in this monograph by Schaid et al. (2), concerns the need to adjust analyses for personal covariates, such as genotypes at other known predisposing genes, or lifestyle characteristics, such as diet or tobacco consumption. The issue of which covariates to include has implications for data collection procedures. Also, there is a need to assess effect modification of lifestyle or environmental influences with the presence or absence of certain genotypes.

**Allelic Loss and Differential Gene Expression Studies**

In contrast to linkage and association studies that evaluate genetic characteristics in diseased and disease-free individuals, allelic loss and differential gene expression studies identify cancer-susceptibility genes by comparing genetic characteristics of cancer cells and normal cells from the same individual. Allelic loss studies identify genes whose proteins prevent a cell from becoming malignant. These genes, the tumor suppressor genes, operate either by repairing damage to a cell’s DNA or, if repair is impossible, by initiating apoptosis. The gene p53 on chromosome 17p is an example of a tumor suppressor gene. A cell that has lost both functioning alleles of p53 (by mutation or deletion) is at high risk of malignant transformation. A cancer composed of the malignant progeny of such a cell provides an ideal testing ground for finding new genes. If the normal cells in the tissue contain two alleles for markers in the region, whereas the cancer cells have lost one or both alleles, there may be a nearby tumor suppressor gene whose loss is responsible for the cancer. A limitation of these allelic loss studies is that the loss in the cancer cells may merely be a consequence of, rather than a cause of, the malignancy.

Like allelic loss studies, differential gene expression studies compare cancer cells with normal cells. However, the objective is not only to identify tumor suppressor genes but also to identify oncogenes whose inappropriate expression is involved in malignant transformation. Genes detectable in the messenger RNA of cancer cells but absent in that of normal cells in the same tissue are good candidates for oncogenes. Comparative genomic hybridization is but one of many techniques recently developed to evaluate gain or loss of genetic material in cancer cells compared
with normal ones. The Cancer Genome Anatomy Project (CGAP) of the National Cancer Institute aims to provide a public database of differential gene expression in cancer cells and in normal cells.

The likelihood of identifying disease genes is greatest for monogenic diseases, such as cystic fibrosis, or for diseases caused by rare, highly penetrant mutations of a gene, such as breast cancer caused by mutations of BRCA1. More problematic are the complex diseases, such as non-insulin-dependent diabetes mellitus (NIDDM) or autism, most likely caused by mutations of multiple low-penetrance genes. For instance, family and twin studies indicate a strong genetic component in NIDDM, and mathematic modeling suggests the interaction of multiple genes. Most studies in NIDDM have relied on the candidate gene approach that has successfully implicated mutations in the glucokinase gene (13) and the glucagon receptor gene (14). As positional cloning becomes more sensitive and powerful, it will complement the candidate gene approach to identifying new genes for complex diseases, although it seems unlikely to supplant it. It also may be necessary to take into account interactions, either among genes or between genes and environmental exposures, to successfully detect additional disease-susceptibility genes.

An important question is when to stop searching for new genes for a specific disease. For a rare disease with a clear mendelian inheritance pattern, such as cystic fibrosis, the answer is straightforward. This inherited disorder, affecting one in 2000 Caucasians, is caused by impaired ion transport because of malfunction of a protein called cystic fibrosis transmembrane conductance regulator. When the gene encoding this protein was cloned in 1989, it was discovered that its mutations cause virtually all cases of the disease. In contrast to the cystic fibrosis story, it is not clear when to stop searching for new genes predisposing to chronic diseases with complex genetic etiologies. For example, it is not known how many distinct genes have polymorphisms with a detectable influence on breast cancer risk. Several investigations (15–17) describe families with multiple cases of breast cancer that are segregating mutations of neither BRCA1 nor BRCA2. As another example, it is not known whether all familial aggregation of ovarian cancer can be explained by mutations of BRCA1, BRCA2, and chance. New analytic methods are needed to address this issue.

Characterizing Attributes of Known Genes

Once a gene that affects disease risk has been identified, there are several analytic tasks at hand. These include estimation of (a) risks associated with specific alleles, (b) frequencies of high-risk alleles in various populations, (c) gene–gene and gene–environment interactions, and (d) the proportion of the disease burden attributable to polymorphisms of the gene.

As the field of genetic epidemiology burgeons and the number of candidate genes increases, it will be increasingly important to give careful consideration to the rationale for investigating a specific gene mutation or allelic variant. Several possible rationales exist: (a) The genetic variant has been associated with familial forms of the disease; (b) the gene codes for proteins involved in biochemical mechanisms of disease pathogenesis; (c) the gene codes for xenobiotic enzymes that are thought to interact with environmental exposures to affect disease risk; and (d) the biologic role of the genetic variant in disease etiology is unclear, but the variant is suspected to be disease associated or has been associated with disease risk in previous studies.

Of primary importance is the strength of the rationale for investigating a given mutation or polymorphic variant. Investigations that are based on rationale (a) are clearly worthwhile because there is strong biologic plausibility that the genetic variant is causally related to disease risk. One example is the discovery of BRCA1 mutations in breast cancer. Following the initial discovery of the BRCA1 gene, subsequent investigations (18,19) sought to estimate the proportion of breast cancer cases that are attributable to BRCA1 mutations by estimating carrier frequencies among familial and sporadic breast cancer cases and among population controls. Rationale (b) also provides a strong basis for investigation if a gene codes for a protein that is known to be involved in disease pathogenesis, as is the case for the amyloid-beta precursor protein in Alzheimer’s disease (20). Investigations based on rationale (c) may be biologically plausible because xenobiotic enzymes involved in the phase I or II metabolism of chemicals may adversely affect the successful detoxification of environmental carcinogens (21). Genetic variants of these enzymes may have deleterious effects by promoting the rapid metabolism of a compound to its toxic metabolite, or it may confer protective effects by successfully metabolizing the compound to a nontoxic form. Examples are the cytochrome (CYP) P450 enzymes responsible for phase I metabolism of environmental chemicals (e.g., debrisoquine hydroxylase [CYP2D6] and risk of lung cancer) (22) and enzymes such as glutathione-S-transferase (GST) that are involved in phase II metabolism (e.g., GSTM1 null genotypes and risk of bladder cancer) (23). However, the reasoning regarding a causal role for a given xenobiotic enzyme is rarely straightforward because of the existence of multiple enzymatic pathways for a specific compound and the fact that some xenobiotic enzymes can be induced by other environmental exposures.

Rationale (d) is the weakest basis for pursuing the study of a putative susceptibility gene and may be responsible for a burgeoning number of inconsistent findings across studies. Several common flaws in study design can contribute to inconsistent results. The first investigations of a putative susceptibility gene may be based on poorly defined or “convenience” controls or may use allelic frequencies for controls that are based on published literature. As a result, findings from early studies can be extremely variable, and resolution requires more definitive studies with the use of appropriate population controls. A related concern is the failure to collect data or adjust for differences in the racial or ethnic distributions of the case and control groups, as discussed in the previous section. Some studies (24) have relied on prevalent cases, which does not allow differentiation of the genetic variants that are associated with better survival (overrepresented among prevalent cases) from variants that increase or decrease the risk of developing the disorder. Characteristics of gene polymorphisms may also affect study feasibility. The population frequencies of the allelic variants may be very low, making it difficult to obtain a sample size large enough to obtain precise estimates of allele frequencies, to estimate main effects, and to estimate gene–gene and gene–environment interactions. Careful consideration must also be given to whether the allelic variants lie in the intron (noncoding) or exon (coding) regions of the gene. If a polymorphism is in an exon (or a close-in promoter region), it is strongly plausible that allelic variants could influence gene expression, and studies addressing the known functional effects of such variants should be consulted. If the allelic variants lie in an intron, they are less likely to have a direct
functional effect but may be in linkage disequilibrium with the true disease susceptibility locus.

In planning investigations of candidate genes, investigators may want to consult an Internet-based resource (Human Epidemiology Genome Network, or HuGENet) offered by the Centers for Disease Control and Prevention (CDC) for on-line publishing of systematic reviews of genetic variants and their associations with the risk of specific diseases. HuGENet is a developing CDC resource for on-line publishing of systematic reviews of certain genetic variants and their associations with the risk of a specific disease. The intent of each HuGENet review is to provide comprehensive background information on molecular or genetic techniques for allelic typing, the prevalence of allelic variants in different racial or ethnic populations, population-based disease risk information, evidence for gene–gene effects or gene–environment interaction, and implications for clinical practice and genetic counseling. Although all of the HuGENet reviews will be peer-reviewed and initially published in peer-reviewed journals such as American Journal of Epidemiology, each on-line review will be updated over time as new information accumulates. The website address for this resource is (http://www.cdc.gov/genetics/hugenet).

**Risks Associated With Specific Alleles**

Estimates of the age-specific and lifetime risks of disease in carriers of specific mutations or polymorphisms are essential for informed clinical management of those individuals with an inherited disease susceptibility. In principle, cohort studies of identified mutation carriers can be used to estimate their age-specific and lifetime risks, but such studies face many challenges. Unbiased estimates of risk require registries with data on a random sample of carriers. Most existing high-risk families recruited for gene identification studies have been selected for the existence of multiple cancers; therefore, their cancer experience tends to overestimate risk in the general population. In addition, because risk is likely to be influenced by both environmental factors and subsequent interventions, observational studies must gather extensive data on baseline risk factors as well as longitudinal measurements on interventions or changing lifestyles. The pros and cons of classic epidemiologic designs for gene characterization are discussed in more detail by Langholz et al. (25) in this monograph. In addition, the family-cohort (also called kin-cohort) design, as described by Struwing et al. (26) and Wacholder et al. (27), has been used to assess the penetrances of mutations. This design is discussed in more detail in the paper by Gail et al. (28) in this monograph.

**Frequencies of High-Risk Alleles**

Following the identification of a candidate gene and its association with a given disease, the next step is to obtain accurate estimates of the frequency of high-risk alleles in the population. This information is important for estimating the fraction of the disease that is attributable to high-risk alleles (see attributable fraction section below) and for determining sample size requirements for subsequent association studies. For gene polymorphisms that are low in frequency (i.e., <10%), large population samples will be needed to obtain precise estimates of the frequencies of high-risk alleles. Furthermore, allele frequencies often vary according to racial or ethnic ancestry, making it very important to collect detailed ancestry information from study subjects so that race-specific allele frequencies can be estimated. In association studies, large sample sizes will be required to estimate the main effects of low-prevalence polymorphisms and to examine gene–gene or gene–environment interaction.

**Gene–Gene and Gene–Environment Interactions**

When several genes have been implicated in a disease and when the frequencies of double mutant carriers are not negligible, we need to know how the genes interact. At present, for example, little is known about possible interactions between the glucokinase and glucagen receptor genes in NIDDM. And, although an interaction in ovarian cancer has been reported between the genes BRCA1 and a polymorphism close to the H-RAS oncogene (29), little is known about the nature of this interaction, if indeed it is found in other independent data. A more complex gene–gene–environment interaction has been observed in research on bladder cancer. Taylor et al. (30) reported evidence of a statistically significant gene–gene–smoking interaction in which the combination with the genotype for slow acetylator status of the N-acetyltransferase (NAT2) enzyme and the presence of one or more NAT1*10 alleles increased the risk for bladder cancer among smokers but not among nonsmokers.

The possibility that modifiable lifestyle characteristics influence gene expression provides hope for the development of preventive strategies. For example, the available data suggest that oral contraceptive use for several years before menopause reduces risk of the type of malignancy that occurs in patients with ovarian cancer without a family history of the disease. The level of protection among carriers of mutations of the BRCA genes is an important issue. The first available data on this issue suggest that oral contraceptive use also reduces the risk of ovarian cancer in carriers of BRCA mutations (31), although its effect on their risk for breast cancer is more problematic (32). Another example is provided by the report in this monograph by Le Marchand (12) on gene–diet interaction in Japanese Americans.

Investigators are beginning to evaluate the relative efficiencies of the various study designs for the detection of interaction among known genes and between known genes and endogenous and exogenous characteristics. Reviews of this issue can be found in three articles (33–35) and in papers in this monograph (6,36). It is well known that large sample sizes are needed for adequate power to detect interactions among environmental risk factors, and this size requirement is also true for gene–gene and gene–environment interaction.

**Attributable Risks**

The public health relevance of a given gene polymorphism is addressed by estimating the proportion of diseased individuals in the population that could be prevented if the high-risk alleles were absent (known as attributable fraction, etiologic fraction, or population attributable risk percent). Accurate estimation of the population frequency of the high-risk genotype is important because attributable fraction is a function of the frequency of the high-risk genotype in the population and the penetrances of the high- and low-risk genotypes. Also, because all of these factors may vary with age, it is important to compute age-specific estimates of attributable fraction. Attributable fractions can also be used to estimate the proportion of disease that is a result of the interaction of a genetic variant and an environmental exposure. Because genotypes are not usually modifiable, the prevention of
disease will depend on interventions that target environmental factors that interact with genetic susceptibility to influence the risk of disease. Study designs for assessing interventions among carriers of high-risk genotypes are the topic of the next section.

Evaluating Disease Prevention Strategies in the Genetically Susceptible

As more genes with predisposing alleles are identified and as people become increasingly aware of genetic developments and interested in knowing about their own genes, there is need to offer them options for preventing the diseases to which they are particularly susceptible by inheritance. To achieve this goal, we must understand the functions of proteins encoded by the normal alleles of the disease-related genes, so that we can devise ways to compensate for these proteins when they are dysfunctional and to eliminate as much of the guesswork as possible in devising and evaluating new interventions. The task of evaluating new interventions presents many design challenges, which are beyond the scope of this paper. The following is a brief discussion of issues in the use of two designs for evaluation: observational cohort studies and randomized trials.

Observational Cohort Studies

It is unlikely that observational cohort studies of carriers can be used to measure the relative benefits of two interventions. Bias can arise in an observational study of interventions when patients at higher risk select more aggressive interventions, as might be the case, for example, if patients with BRCA1 mutations and with a high incidence of familial breast cancer tend to select prophylactic mastectomy more often. Bias also can arise if large treatment centers that see a disproportionate share of high-risk patients offer only a single intervention option to identified carriers.

Randomized Trials

The feasibility of randomized trials to compare interventions depends strongly on the disease and on the estimated risk of study participants. The problem of conducting a randomized trial comparing interventions in breast cancer illustrates some of the issues. It seems unlikely that a randomized trial comparing prophylactic mastectomy with a nonsurgical intervention could be completed successfully. Indeed, any such trial will need two active arms, and it cannot contain a no-treatment control. Presently, there is not widespread agreement on a potentially effective nonsurgical intervention for young, high-risk women. There are, however, more options for randomized trials of interventions in carriers of the genes for hereditary nonpolyposis colon cancer. There is a strong rationale for nonsurgical interventions such as nonsteroidal anti-inflammatory drugs (e.g., aspirin) and dietary supplements with calcium or folic acid. Moreover, colonoscopy followed by sampling of the mucosal wall of the colon can be used for the periodic evaluation of study subjects (37).

The recently completed tamoxifen-versus-placebo and the ongoing raloxifene-versus-tamoxifen breast cancer prevention trials select women at high risk of breast cancer and thus contain higher prevalences of BRCA1 and BRCA2 mutation carriers than in the general population. Still, it is not clear that these trials contain sufficient numbers of carriers to evaluate the efficacy of these interventions among carriers.

Practical Constraints

The practical constraints faced by the genetic epidemiologist include those that impede classic epidemiologic research on the etiology of environmental exposures in chronic diseases. These constraints include the need to minimize selection bias and response bias, to ensure accurate and complete data collection and appropriate statistical analysis, and to interact responsibly with all study participants. However, the genetic epidemiologist also faces constraints special to genetic research. These constraints include problems particular to family studies and family registries as well as problems particular to certain ethnic groups.

The study of cancer risk, cancer risk factors, and intervention efficacy in carriers of disease-susceptibility genes is challenging in several ways. Major challenges arise from the ethical issues in connection with the high cancer risks of study subjects and their offspring, but there are other logistical issues as well. Because the prevalence of carriers of rare mutations is small, large, multisite, expensive studies must be mounted. Cohort studies of carriers and their relatives must be based on random samples of carriers. Retrospective cohort studies of cancer incidence following various interventions in high-risk families must avoid the selection bias of counting those cancers that brought the family to attention. Finally, randomized trials must compare two active treatment arms because allocating an inactive placebo to very high-risk individuals is unethical.

Response Rates and Selection Bias

The problems of incomplete response that plague all epidemiologic studies apply to genetic studies as well, particularly to those attempting to characterize genetic polymorphisms and gene–environment interactions. If response rates differ by ethnicity, comparisons of gene frequency across groups are compromised. Moreover, family-based studies are vulnerable to selection bias by probands with a family history of the disease under study. Potential participants who have reason to believe that they may carry high-risk alleles may refuse to provide DNA because of concerns of employment or medical discrimination for either themselves or their relatives.

Problems Particular to Family Registries

Individuals with cancer sometimes prefer that their relatives not know their disease diagnosis. In addition, genetic testing of multiple family members can raise difficult ethical problems. For example, certain members of multiple-case breast cancer families may prefer not to know their genetic status with respect to BRCA1, but it may be clear from a positive test result in an offspring. More detailed discussion of the issues special to family registries can be found in this monograph in the papers on a breast cancer registry by Hopper et al. (38) and on a colorectal cancer registry by Haile et al. (37).

Problems Particular to Certain Ethnic Groups

Members of different ethnic groups have distinct cultural patterns for accepting and handling information on chronic diseases (39). These patterns often restrict the amount of information shared among family members, and they must be accommodated in attempting to gather family histories of disease. A further complication is the likelihood that first-generation immigrants often do not know important details of their relatives’ medical
history. In addition, there are difficulties in obtaining interview data and biologic specimens from relatives in other countries. Furthermore, we have found ethnic citizens of the United States of Asian ancestry to be less willing to give blood samples than individuals from other groups.

**Ethical, Psychosocial, and Legal Issues**

Each of the study designs described above requires informed written consent from participants, but there is uncertainty regarding the proper use of consent in the genetic setting. Two

---

**Table 1. Study designs for identifying and characterizing disease-susceptibility genes**

<table>
<thead>
<tr>
<th>Goal</th>
<th>Design</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identify new disease genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linkage studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parametric linkage</td>
<td>Good power when assumptions agree with mechanism generating data</td>
<td>Requires knowledge of mode of inheritance; can be sensitive to violations of assumptions</td>
<td></td>
</tr>
<tr>
<td>Nonparametric linkage</td>
<td>Requires no assumptions about mode of inheritance, marker location</td>
<td>Impaired by presence of phenocopies, etiologic heterogeneity</td>
<td></td>
</tr>
<tr>
<td>Association studies</td>
<td>Statistically efficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case-unrelated control</td>
<td>Straightforward collection of cases and unrelated controls</td>
<td>Confounding by ethnic stratification</td>
<td></td>
</tr>
<tr>
<td>Case-related control</td>
<td>Controls for ethnic ancestry</td>
<td>Less powerful than linkage studies for rare, high-penetrance alleles</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Characterize known disease genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate allele frequency</td>
<td>Cross-sectional studies</td>
<td>Easy to conduct</td>
<td>Genetic variants associated with poor survival underrepresented in cross-sectional sample</td>
</tr>
<tr>
<td>Estimating penetrance</td>
<td>Case-unrelated control</td>
<td>Good power for examining common, low-penetrance polymorphisms</td>
<td>Penetrance may be difficult to assess because mild phenotypes may be undetected in control subjects</td>
</tr>
<tr>
<td></td>
<td>Family cohort</td>
<td>Effective design for assessing penetrance of rare, high-penetrance mutations</td>
<td>Incomplete penetrance can affect measures of familial aggregation</td>
</tr>
<tr>
<td>Evaluating environmental modifiers of penetrance</td>
<td>Case-unrelated control</td>
<td>Efficiency is good for examining common polymorphisms Acknowledges that not everyone with genotype will manifest phenotype</td>
<td>Requires knowledge of mode of inheritance Can be biased because of selection factors for probands</td>
</tr>
<tr>
<td></td>
<td>Case-related control</td>
<td>Controls for ethnic ancestry</td>
<td>Confounding by ethnic stratification</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Evaluate preventive strategies in genetically susceptible</strong></td>
<td>Randomized trial</td>
<td>Usually free of selection, confounding biases Most valid study design for assessing effect of intervention</td>
<td>Feasibility and power depend on number of carriers, size of risk among genetically susceptible individuals</td>
</tr>
<tr>
<td></td>
<td>Cohort study</td>
<td>Retrospective cohort studies can be economical</td>
<td>Patients at high risk may selectively seek aggressive interventions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Patients at high risk may selectively seek aggressive interventions</td>
</tr>
</tbody>
</table>

---

Journal of the National Cancer Institute Monographs No. 26, 1999
issues have been the subject of considerable recent debate. First, must donors give additional written informed consent for the use of archived blood or tissue specimens collected in studies whose purposes differ from those currently planned? The National Bioethics Advisory Commission, a group of lawyers, ethicists, and medical professionals, has recently issued a draft report arguing for tighter controls on stored materials to protect the donors’ privacy (posted on the Internet at www.bioethics.gov). Specifically, new informed consent may be necessary whenever the investigators have not indicated previously to participants the possibility that samples may be used for purposes other than the original study objectives (40). Moreover, donors at this point should be given a chance to ask that their specimens not be used in this or any future research project. If a study involving de novo acquisition of specimens might pose more than a minimal risk of harm to the donors, Institutional Review Boards must ensure that the risks are clearly described to donors before consent is sought. In addition, all potential donors should be informed that their biologic materials will be stored for years and that their specimens may subsequently be reanalyzed for currently undetermined research purposes (41).

Second, there has been uncertainty concerning the responsibility of the investigator to notify study participants if genetic testing indicates the presence of a genetic abnormality that confers increased cancer risk. The current view on this issue is that, in obtaining consent for participation in new research, the investigator should indicate exactly what information participants will receive and when this information will be available, i.e., the investigator need not be obligated to reveal genetic findings to participants, but his or her plans to withhold such information must be delineated clearly at the outset (42,43). According to the National Advisory Council for Human Genome Research (44), the decision whether or not to notify participants should depend on the accuracy with which the data and the test predict risk; the efficacy of existing prevention measures; the availability of non-directive education and counseling for family members; and the likelihood of genetic discrimination with respect to health insurance, life insurance, and employment opportunities.

**CONCLUSIONS**

Table 1 summarizes the various designs available to the genetic epidemiologist for gene identification and characterization as well as some of the strengths and weaknesses of these designs. As evident from the table, each design has its drawbacks, and no one design is clearly preferable for achieving a given goal. The classic case–control study based on unrelated individuals with and without the disease of interest tends to be more efficient than family-based, case–control designs that inherently match case patients and control subjects on ethnic ancestry and may also match them on environmental factors. The classic case–control studies also tend to be simpler to conduct and analyze. There is a need for more research on the magnitude of the potential problem of confounding by ethnic admixture that compromises inferences from these studies. This problem can be alleviated somewhat by focusing on polymorphisms with functional significance in the etiology of the disease, combined with statistical methods that adjust for ethnic stratification in the analysis phase of the study.

The workshop included lively discussion of the desirability and feasibility of designing population-based studies to meet all or most of the goals involved in identifying and characterizing genes. In this monograph, Zhao et al. (45) present an integrated approach to gene identification and characterization involving ascertainment of families via individuals with cancer (probands) identified in population-based cancer registries. This approach is illustrated here by the population-based family registries for breast cancer (38) and colorectal cancer (37,46). When this ascertainment scheme is used, one must decide at the onset on priorities for the extensive task of recruiting extended families, to verify reported disease history, to gather personal data, and to obtain biologic specimens.

Despite the ethical and strategic difficulties in conducting them, adequately controlled observational and randomized studies provide the best mechanisms for progress in evaluating the genotypes predisposing to increased cancer risk, the prevalences of these genotypes, and the lifetime risks of site-specific cancers borne by carriers of these genotypes. It is important that research protocols actively recruit ethnic minorities and, when appropriate, members of both sexes to estimate possible gene–ethnicity and gene–sex interactions.

**REFERENCES**

Mutations in BRCA1 and BRCA2 in breast cancer families: are there more breast cancer susceptibility genes? Am J Hum Genet 1997;60:486–95.


(37) Haile RW, Siegmund KD, Gauderman WJ, Thomas DC. Study-design issues in the development of the University of Southern California Consortium’s Colorectal Cancer Family Registry. Monogr Natl Cancer Inst 1999;26:89–93.


NOTES

Supported by Public Health Services grants 5R35CA47448 to Dr. Whittemore (National Cancer Institute), 3RO1NS31964 to Dr. Nelson (National Institute of Neurological Disorders and Stroke), and 5RO1ES08150 to Dr. Nelson (National Institutes of Environmental Health Sciences), National Institutes of Health, Department of Health and Human Services.

We are grateful to Dr. Beth Newman for discussions that helped in the preparation of this paper.