Integrated Designs for Gene Discovery and Characterization

Lue Ping Zhao, Corinne Aragaki, Li Hsu, John Potter, Robert Elston, Kathleen E. Malone, Janet R. Daling, Ross Prentice

Recent advances, including near completion of the human genome map, ever improving high-throughput technologies, and successes in discovering chronic disease-related genes, have stimulated the further development of genetic epidemiology. The primary mission of genetic epidemiology is to discover and characterize genes, whether independent of or interactive with environmental factors, that cause human diseases. To accomplish such a mission, genetic epidemiology needs to integrate both genetic and epidemiologic approaches. One of the challenges facing such an integrated approach is the identification of study designs that are efficient for both gene discovery and characterization. Because designs for gene discovery alone and designs for gene characterization alone have been elaborated in the other two panels, the focus of this paper is to describe those designs that may be useful for discovery and characterization jointly, including case-family and case-control-family designs. Examples of integrated designs are described, and studies of breast cancer conducted at the Fred Hutchinson Cancer Research Center are used for illustration. Finally, related analytic issues are also discussed. [Monogr Natl Cancer Inst 1999;26:71–80]

The ultimate goals of genetic epidemiology include not only the discovery of novel functional genes and investigation of their functional properties—the primary objective of gene discovery—but also the establishment of population allele frequencies and genotype penetrances in relationship to specific human diseases—the primary objective of gene characterization. Complexities of common human diseases, from a genetic perspective, have been described by Schaid in the Gene Discovery Panel (1), earlier by Lander and Schork (2), and by others. These complexities make research more difficult not only in gene discovery but also in gene characterization. To succeed in dissecting complex diseases, one needs to consider a comprehensive approach, based on the analysis of systematically collected data with the use of multiple analytic tools. Earlier in 1997, we described a design framework (3) to integrate population-based and family-based designs using a multistage approach (4–8). As a framework, it is inclusive of most study designs used to discover genes, such as affected sib-pairs and highly selected families, as well as of designs to characterize candidate genes, such as population-based case-control and case-relative-control study designs. Because of its inclusive nature, one utility of this framework is to serve as a paradigm for categorizing different study designs.

Because design topics on gene discovery and on gene characterization have been explored fully by investigators in the Gene Discovery Panel and Gene Characterization Panel, respectively, this paper will focus on integrated designs for the purpose of both gene discovery and gene characterization. After describing the need for and challenges to integration, this paper briefly introduces the population-based family-study design framework, including a description of several new results beyond those discussed in our earlier paper (3). We list examples of designs for both gene discovery and gene characterization and then attempt to evaluate critically this framework from various perspectives. For illustration, we describe studies of breast cancer conducted at the Fred Hutchinson Cancer Research Center. Finally, we outline a likelihood framework that may be useful for developing statistical methods, taking advantage of this design framework, for calculating sample sizes and study power, for comparing methods, and for developing robust and efficient methods.

RATIONALE

The majority of common diseases, especially cancer and coronary heart disease, appears to have a complex etiology, including incomplete penetrance, phenocopies, genetic heterogeneity, gene–gene interactions (epistatic), and gene–environment interactions. Dissecting complex traits calls for an interdisciplinary approach. Recognizing this emerging need, researchers with diverse scientific backgrounds have begun interdisciplinary communication and, in some cases, have collaboratively proposed interdisciplinary studies.

In such research, one of the first challenges is how to design such a study that is efficient and appropriate from both genetic and epidemiologic perspectives. From a genetic perspective, studies should focus on informative samples, improving the chance of finding disease genes. Such considerations often lead geneticists to select families with "unusual" family histories, since such families tend to have an enhanced probability of carrying disease genes. From the perspective of epidemiologists, population-based sampling is key to reliable detection of significant etiologic factors in the study population. Furthermore, under population-based sampling, research results can be generalized to the target population. Motivated by such considerations, epidemiologists tend to design studies that systematically identify case patients and control subjects in population-based, case-control studies. Cohort studies are also motivated by the same considerations but, to keep the discussion focused, will not be elaborated here. Unfortunately, these genetic and epidemiologic

Affiliations of authors: L. P. Zhao, C. Aragaki, L. Hsu, J. Potter, K. E. Malone, J. R. Daling, R. Prentice, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA; R. Elston, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH.

Correspondence to: Lue Ping Zhao, Ph.D., Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave., N, MW-806, Seattle, WA 98109.

See "Notes" following "References."

© Oxford University Press
perspectives appear to be at extreme ends of the research spectrum, and their integration has thus been a challenge to the development of interdisciplinary work, motivating the development of the population-based family study design framework to be introduced below.

Bearing in mind the genetic and epidemiologic perspectives, one cannot assess a particular study-design choice without appropriate consideration of the corresponding statistical analyses. Table 1 lists some typical statistical analyses performed in epidemiologic and genetic studies. Association analysis, primarily used in epidemiology, assesses association of the disease phenotype with candidate genes as well as with environmental factors. To establish initial evidence to motivate a genetic endeavor, one may conduct an aggregation analysis to examine 1) presence of familial aggregation of disease and, if established, 2) the patterns of familial aggregation as well as 3) evidence in support of any genetic hypothesis. To confirm a genetic hypothesis, one can perform a segregation analysis, providing a tentative estimate for the penetrance and allele frequency of a postulated disease-related gene. Success in these initial analyses leads to a mapping study, via linkage analysis and linkage-disequilibrium analysis. Whereas the goal of linkage and linkage-disequilibrium analysis is to discover disease loci by studying inherited mutations, loss-of-heterozygosity analysis using tumor tissues can be used to study somatic changes in solid tumors, i.e., to search for genes that may be responsible for initiation, promotion, and progression of solid tumors. Many of these analyses may be integrated into a combined analysis to gain efficiency, as described below.

An immediate benefit of having an integrated design is that the resultant study not only facilitates multiple analyses but may also lead to a gain in statistical efficiency. For example, when linkage and linkage-disequilibrium approaches are combined, the resulting analysis may allow one to detect the position of putative disease genes with improved efficiency. One may also combine linkage-disequilibrium with association analysis so that one can adjust for known etiologic factors and can search for putative factors that independently or interactively contribute to disease. Other benefits of having integrated designs from a statistical perspective are enumerated in the “Methodologic Considerations” section.

**Brief Introduction to Population-Based Family Study Designs**

We have previously described a general framework for designing population-based family studies (3). Although the framework is general (see “Examples of Integrated Design” section), a comprehensive design would have three sampling stages: association and aggregation (A), segregation (S), and linkage (L), as illustrated in Fig. 1. At stage A, a study may adopt, e.g., a case–control design, sampling a group of case patients within a specific time period from a well-defined population, e.g., a population covered by a cancer registry, or the enrollees of a health maintenance organization, and also sampling a group of unaffected individuals from the appropriate comparison population. From each case patient or control subject, researchers gather questionnaire-based information and biologic samples systematically. The primary objective at stage A is assessment of the associations of candidate genes and environmental factors with a disease phenotype, with the secondary objective of assessing familial aggregation of the disease phenotype with the use of family history data. At stage S, the study ascertains pre-specified relatives of case patients and control subjects, gathers questionnaire-based information from all participating relatives, and collects biologic samples from a subset of participating relatives. The primary objectives at stage S include (a) assessing residual familial aggregation after adjusting for candidate genes and environmental factors as covariates and assessing whether the aggregation pattern supports a genetic hypothesis, then (b) describing the penetrance of the putative gene and corresponding allele frequency via segregation analysis, and, finally, (c) quantifying the residual familial aggregation after adjusting for the putative gene and those covariates. At stage L, the researcher selects and determines markers in families that appear to have unusual disease patterns, since such families and pertinent family members are likely to carry abnormal alleles. In addition, stage L may include relatives who have not participated in stage S. To take advantage of high-risk family registries, stage L can even include new families that are from the same general population. The primary objective of stage L is to localize genes, which, independently of or interactively with candidate genes or environmental factors, cause the disease phenotype.

This framework (3) has many features noted earlier, some of which are briefly described here. As a framework, it encompasses many designs that are suitable for either discovering genes or for characterizing candidate genes. Combining these two types of designs leads to the development of “hybrid designs” that can be used for both gene discovery and gene characterization. This melding of designs is made possible by multistage sampling. Although each stage has its own primary objectives, data collected from all stages can be analyzed jointly, and the findings are then interpretable in the context of the underlying general population.

While recognizing desirable features of this framework, one

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Symbol</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association analysis</td>
<td>A</td>
<td>Testing a direct association of observable covariates with the phenotype</td>
</tr>
<tr>
<td>Aggregation analysis</td>
<td>Agg</td>
<td>Establishing familial aggregation of the phenotype</td>
</tr>
<tr>
<td>Segregation analysis</td>
<td>S</td>
<td>Evaluating possible penetrance and allele frequency for a putative disease gene via a random effect model</td>
</tr>
<tr>
<td>Linkage analysis</td>
<td>L</td>
<td>Testing the presence of cosegregation of markers with putative disease genes</td>
</tr>
<tr>
<td>Linkage-disequilibrium analysis</td>
<td>LD</td>
<td>Testing allelic association between markers and putative disease genes</td>
</tr>
<tr>
<td>Loss-of-heterozygosity analysis</td>
<td>LOH</td>
<td>Examining genomic instability in abnormal tissues</td>
</tr>
<tr>
<td>Combined analysis</td>
<td></td>
<td>Combining two or more of the above analyses into one for possible gain in efficiency, compared with separate analysis</td>
</tr>
</tbody>
</table>
needs to also bear in mind its limitations. Probably the most significant limitation is that following such a design framework through multiple stages may complicate study conduct and may even encumber studies of gene discovery. To avoid delays in gene discovery, it is necessary to design a protocol that facilitates the quick identification of high-risk families without invalidating the overall study design.

Since our publication on this design framework, we have found some additional related features of interest. First, this framework may be considered as an extension of the epidemiologic design paradigm by integrating family studies into population-based studies (Aragaki, personal communication). This recognition helps summarize the framework in a succinct manner, and it may help epidemiologists appreciate the complexity of conducting population-based family studies. Second, a three-stage study may allow one to discover new genes, to characterize genes, and again to discover additional genes after adjustment for discovered genes, forming a cycle of gene discovery and characterization (see Fig. 2). Theoretically, such a cycle allows researchers to uncover more than one gene at a time and poten-

Fig. 1. Flowchart for a three-stage population-based family study design. Subscript $p$ = proband; subscript $r$ = relatives; subscript $a$ = additional relatives; subscript $n$ = new families, lower case letters = a scalar variable; upper case letters = a vector; $D$ or $d$ = disease phenotypes; $E$ or $e$ = environmental factors; $B$ or $b$ = biologic samples; $M$ or $m$ = markers; and $FH$ or $fh$ = self-reported family history.

Fig. 2. Schematic flowchart for a three-stage population-based family study. Lower case letters for individual data, upper case letters for all data in families; subscript $p$ = proband; subscript $r$ = relative; $d$ = the disease state; $e$ = environmental covariates; $b$ = biologic samples; $fh$ = self-reported family history; and $m$ = genetic markers.
tially to discover many genetic factors in a single study. Although conceptually appealing, this approach will likely experience limited power at some point of the cycling process, when the majority of the familial aggregation has been explained or when residual familial aggregation is caused by poorly measured environmental factors. Third, we have taken a more detailed look at linkage-disequilibrium analysis and its potential for mapping complex traits. Our experience to date, based on simulated genome scan data (Aragaki, personal communication) (9), supports the idea that the combined analysis of both linkage and disequilibrium signals should be a powerful tool for fine-scale mapping. Unlike linkage analysis, linkage-disequilibrium analysis is preferably carried out on individuals or families that are systematically ascertained from a well-defined population to minimize potential biases as a result of ascertainment or population admixture. In this sense, design considerations for gene discovery and for gene characterization are somewhat convergent. Fourth, the development of a high-density genome map also facilitates a genome scan for loss of heterozygosity. To have an accurate view of the genome-wide pattern of loss of heterozygosity, one generally requires systematic collection of normal and abnormal tissues at various disease stages, calling for special design consideration within this framework.

**Examples of Integrated Designs**

**Case-Only Study**

The case-only study design involves sampling population-based cases (or systematically ascertained cases from a well-defined cohort) and may be considered as a study with a single stage A. From each case patient, a tissue specimen is taken, and its genomic DNA is extracted and genotyped for an array of genetic markers. Under typical population-genetic assumptions for the study population, genetic markers in the population may be thought of as independently distributed and thus follow Hardy–Weinberg equilibrium. However, in the presence of linkage-disequilibrium at certain marker loci, the genetic markers from patients would deviate from Hardy–Weinberg equilibrium proportions. Scanning through the human genome, one can test for association between marker alleles and putative disease-related alleles (10).

When solid-tumor tissue samples are collected from patients, both the abnormal part of the tissue and the adjacent normal tissue may be extracted and genotyped for the same genetic markers. A direct comparison of genetic markers between normal and abnormal tissues would reveal the presence or absence of loss of heterozygosity at those loci. The pattern of loss of heterozygosity is indicative of genomic instability that occurred during disease initiation, promotion, and progression. Combining linkage-disequilibrium and loss of heterozygosity, one may be able to detect those genetic alterations that are important as both germline and somatic mutations. Although this design is useful for the stated objectives, it has some limitations in the assessment of gene–environmental interactions (to be discussed in “Discussion and Summary” section).

**Case-Relative–Control Study**

The case-relative–control study involves sampling population-based case patients and their unaffected relatives as matched control subjects, which again may be thought of as a stage A study. Although biologic samples collected from cases can be used for linkage-disequilibrium and loss-of-heterozygosity analyses as described above, the primary reason for including relative controls is to facilitate an analysis of the association between genetic markers and disease phenotype. Improving on the usual population-based, case–control design, the case-relative–control study is better able to yield unbiased estimates of odds ratios for the association of candidate genes with phenotype because it potentially adjusts for the confounding effects as a result of population admixture or genetic heterogeneity because of different founders (11,12) and for shared environmental factors. While appreciating the usefulness of the linkage-disequilibrium analysis, we need to realize that this design, and those described later (two-stage case-family–control and two-stage case–control family studies), are not efficient for linkage analysis. Such a design is justified primarily on the ground of cost efficiency and feasibility for mapping complex traits using association, linkage-disequilibrium, and loss-of-heterozygosity analyses.

**Two-Stage Case-Family Study**

The case-family study involves sampling population-based case patients at stage A and ascertaining family members (prespecified by degree of kinship), regardless of their disease status at stage S. Because of the potentially large number of relatives in such studies, collection of biologic samples often has to be restricted to case patients and, possibly, affected relatives and selected unaffected relatives. For example, a case-family study of a common disease may sample incident case patients from a well-defined population at stage A and then ascertain all their first-degree relatives, including parents, siblings, children, and spouse(s), at stage S. Even for “common” diseases, the majority of case-relatives are unaffected at the time of the study. Hence, one or more unaffected relatives could be used as control subjects, and the corresponding analyses for a case-family–control study may be performed. If all of the relatives are included, however, one may perform two additional analyses: aggregation analysis and segregation analysis. Through aggregation analysis, one can estimate the magnitude of the familial aggregation of phenotypes and can establish patterns of familial resemblance. Furthermore, a combined association and aggregation analysis, after adjusting for known covariates (candidate genes or environmental factors), may identify residual familial aggregation and may indicate whether or not it supports a specific genetic hypothesis. Once a genetic hypothesis is formulated, one can use the combined association and segregation analysis to estimate the penetrance parameters of potential putative gene(s) and the corresponding allele frequencies. Furthermore, a combined association, segregation, and aggregation analysis may be useful in quantifying residual familial aggregation after adjusting for the putative gene(s), in addition to covariates. An absence of residual familial aggregation would suggest limited additional genetic contribution to the complex disease. Although such a case-family dataset is useful for all of these purposes, our experience to date indicates that a combined association and aggregation analysis, or any other combined analysis that includes those two analyses, may experience nonidentifiability of key parameters in the absence of external information. Furthermore, from an epidemiologic perspective, such a study tends to overweight on environmental factors (and candidate genes) that are shared within families and, therefore, has reduced power to study environmental factor-related associations. To address this defi-
ciency, one may include population-based control subjects as described in the two-stage case-family control study. From a genetic perspective, however, limited power for linkage and linkage-disequilibrium analysis would be a serious concern. Addressing such a concern motivates the third stage of the data collection to be described below.

**Three-Stage Case-Family Study**

Following the paradigm outlined above, one may identify those families, satisfying the prespecified sampling criteria, with an “unusual” family history. The study then ascertains all members of those families because those relatives are more likely to carry genetic defects of interest. The purpose at this stage of data collection, known as stage L, is to optimize the efficiency of linkage and linkage-disequilibrium analysis. Also at stage L, one may include additional families that are samples of convenience from, e.g., high-risk family registries, which meet certain sampling criteria and yet have not been part of the first two sampling stages. Although naively pooling such ad hoc samples is inappropriate for association and aggregation analysis, adding such families is valid for linkage analysis, provided that these highly selected families are from the same study population and thus share the same penetrance and allele frequency. This assumption is, unfortunately, not readily verifiable with available data. However, the ultimate proof of a successful linkage analysis is the localization of disease genes.

**Two-Stage Case-Family–Control Study**

Extending the two-stage case-family study design, the two-stage case-family–control study design collects, in addition to those cases and their prespecified relatives, population-based controls at stage A. The primary reason for including population-based control subjects is to improve the efficiency of assessing either environmental and lifestyle variables or candidate genes in relation to disease phenotype and, more important, to improve the chance of finding genes that may be involved in gene–candidate gene interactions or in gene–environment interactions.

Following the same rationale for extending the two-stage case-family study to the three-stage case-family study, a three-stage case-family–control study adds highly selected families at stage L to optimize the efficiency of performing linkage and linkage-disequilibrium analysis.

**Two-Stage Case–Control Family Study**

The two-stage case–control family study extends the two-stage case-family–control study design by including prespecified relatives of control subjects at stage S, a design that has also been detailed by Hopper et al. (13). The usually large number of control relatives necessarily limits collection of biologic samples and questionnaire information. In addition, the prevalence of many human diseases may be low among relatives of control subjects, e.g., from 4% to 7% for breast and colorectal cancers. Despite these limitations, control families are valuable for (a) assessing familial aggregation of environmental factors or lifestyle variables in the general population, which may be compared with the pattern in families of cases and is informative about the role of potential environmental factors, and (b) assessing aggregation and segregation through comparing disease occurrence among relatives of case patients with that among relatives of control subjects, the analytic tools that require fewer assumptions and hence lead to more robust conclusions. In practice, the decision to sample control families has to be made on a case-by-case basis, balancing the information gain against the required data-collection effort and cost.

As described above, one may be interested in adding stage L to the above two-stage case–control family study, resulting in a three-stage case–control family study. In principle, this stage includes high-risk families of case patients, or those of control subjects, or even families from external sources.

**Evaluation of Integrated Designs**

Integrated designs for gene discovery and for gene characterization represent a shift in the research paradigm. In spite of the motivations listed above, such a shift should be critically evaluated. In this section, our objective is to evaluate integrated designs from genetic, epidemiologic, and analytic perspectives by considering the following questions.

Why would one want to integrate designs for gene discovery and characterization? In contrast to integrated designs, a traditional approach is to design a gene discovery study and, if one or more genes are mapped, to design or use a separate population-based study to characterize them. Naturally, geneticists in the gene discovery study could optimize the study efficiency by using, for example, several large pedigrees or a series of affected relative pairs. Conversely, epidemiologists could follow the lead of mapped genes and could design a separate study to characterize the genes. A successful example is the story of discovering and characterizing the BRCA1 gene. The primary advantages of this traditional approach include 1) the simplicity of designing focused research projects, which ensures the timely discovery of disease genes, and 2) the feasibility of conducting such projects in stages, so that the characterization study is not proposed until it is proven necessary. The disadvantages include 1) the difficulty of assessing the importance of discovered genes in the general population, since assessing penetrance and allele frequency is not feasible with high-risk families; 2) the overly long time lag before mapped genes are characterized, therefore slowing down the translational research to clinical practice or to cancer prevention and control; and 3) inefficiency because genotyped data from highly selected families are not easily used for characterizing genes.

An integrated design may overcome these disadvantages. As noted earlier, most designs for gene discovery and for gene characterization are encompassed by this framework, and the actual choice of a particular design depends on the objectives and the available resources. For example, if discovering genes has a high priority, one could design a three-stage study as follows: At stage A, one would systematically ascertain all cases and administer a simple family-history questionnaire, without collecting either detailed information or biologic samples. On the basis of simplified family-history information, one can ascertain relatives of interest at stage S, and a minimum amount of information is collected. Information on relatives, especially verified disease phenotypes, can now be used for sampling at stage L. Extensive information and biologic samples are gathered from all family members and are used for gene discovery. Once promising genes have been identified, one can now contact probands and their relatives gathered at stages A and S for characterizing genes.

Why population based? Whereas population-based designs have been extensively used in epidemiology to ensure unbiased
estimation and valid inference, their value is increasingly appreciated in genetic research. For example, if a linkage-disequilibrium analysis is used to map disease genes, it is essential to ensure that the study samples are population based; e.g., haplotypes with the normal and abnormal alleles should be representative of their respective populations. Population-based designs become even more critical to the estimation of allele frequencies in the general population, an important aspect of gene characterization. Furthermore, if there is an interest in gene–environmental interactions, population-based samples become even more important to ensure valid estimation and inference.

Why family based? The essence of gene discovery is to find putative genes that explain the disease aggregation within families; hence, family-based designs are commonly used in genetic research. In contrast, epidemiologic studies generally focus on the direct association of measured genes or environmental factors with disease phenotype and tend to use, instead of family data, independent case patients and control subjects so that no genetic or environmental factors are overmatched between family members. Nevertheless, these contradictory design considerations have been converging in recent years, as family-based, case-relative–control designs have been advocated for gene characterization studies. The primary reason for using a case-relative–control design is to overcome the confounding effects as a result of population admixture, especially for studies in the multiethnic, multiple founder population like the United States.

Is there a single optimal integrated design? An optimal design choice depends on the nature of the disease phenotypes and the underlying disease genes to be discovered. Furthermore, the choice depends on the resources available to investigators. Nevertheless, being guided by the above design framework, investigators could think through issues relating to gene discovery and characterization and could come up with a desirable integrated design for the specific problems to be addressed.

What are the practical utilities of integrated designs? The above design framework may help investigators think through many critical issues for both gene discovery and gene characterization. One important utility is to set research priorities and to allocate research resources. For example, the Cooperative Family Registries of Breast and Colorectal Cancer as well as the Family Registries of Breast and Colorectal Cancer also represent such long-term resource development activities for cancer genetic research, at least for the next decade. If appropriately designed, such resources could be efficiently used to answer a wide range of research questions. A second utility of this framework is that it facilitates the planning of interdisciplinary research projects, such as program project grants having several components. By using a well-planned research framework, projects within such a program can share all the data being gathered and can perform individual analyses using all available information from all projects to achieve efficiency and population-based interpretation. A third utility is that it helps individual investigators prioritize a long-term research agenda.

**Population-Based Family Study of Breast Cancer**

In 1993, we (Zhao, Malone, Daling, and Ostrander) embarked on a genetic epidemiology study designed to investigate the independent and interactive contributions of genetic and lifestyle or environmental factors to the familial aggregation of breast cancer in a well-defined population-based series of young case patients with breast cancer and age-matched control subjects. The goals of the study include the conduct of aggregation and segregation analyses that account for exposure to risk factors for breast cancer in both probands and their relatives, the identification and characterization of new candidate genes for breast cancer, the screening of case patients and control subjects for identified susceptibility genes for breast cancer, and the assessment of interaction between environmental and genetic factors. As an example of the proposed design framework described above, this study shows the versatility of population-based studies for accommodating diverse scientific goals.

**Original Case–Control Studies**

Two population-based case–control studies of breast cancer in young women provided the foundation for this genetic epidemiology study. Both studies, led by Dr. Janet Daling, were designed to evaluate the role of reproductive factors and oral contraceptives in the etiology of breast cancer and used similar methods. For the first study, all incident cases of breast cancer diagnosed from January 1, 1983, through April 30, 1990, among women born after 1944 who were residents of three western Washington counties were ascertained and approached for interview. Case patients were identified through the Cancer Surveillance System, a population-based cancer registry that participates in the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute. The Cancer Surveillance System ascertains more than 99% of all incident cancer cases in the 13 western Washington counties. A population-based control group was ascertained with the use of random-digit telephone dialing within each of the three counties from which the cases were drawn. A total of 845 women with breast cancer and 961 control subjects were successfully interviewed (83.2% of eligible case patients and 75.5% of eligible control subjects, respectively). In the second case–control study, all incident cases of breast cancer diagnosed from May 1, 1990, through December 31, 1992, in women under the age of 45 years residing in the three-county area were ascertained and approached for interview. Interviews were completed with 643 women with breast cancer and with 610 control subjects (86.4% and 78.1% of those eligible). Both studies used standardized in-person questionnaires that elicited information on a wide array of risk factors, including reproductive, contraceptive, and menstrual history; lifestyle factors, such as alcohol, smoking, and dietary intake; medical history; family history (enumeration of female relatives, questions on years of birth and death, history of cancer, and age at diagnosis for each relative); and demographics.

**Other Ancillary Studies**

Case patients and control subjects identified in these two studies have been used for several additional ancillary studies. Currently, we are following all 1288 previously interviewed women who were diagnosed with invasive breast cancer and who agreed to participate in the current study. The case patients (or their proxies) are approached for completion of a questionnaire eliciting data on exposures after diagnosis, treatment history, and recurrences. To date, questionnaires and release forms for tumor access and chart review have been obtained for more than 86% of the case patients. Tumor tissue blocks are obtained by the project pathologist, Dr. Peggy Porter, from cases for
evaluation of potential prognostic factors, including markers of proliferation, tumor suppressors, cell-cycle genes, and genes involved in apoptosis, and medical records are reviewed to evaluate and control for the impact of treatment.

**Genetic Epidemiology Study**

The previously interviewed case patients and control subjects were recontacted in this study and asked to update and expand previously reported family history information. Female relatives (mothers, sisters, aunts, cousins, grandmothers, or their proxies) of 550 case patients and 550 control subjects (from the second case–control study) were asked to provide data on their own risk factor histories through a telephone interview (5500 interviews completed thus far), and relatives from selected families, such as those with four or more affected members, were asked to provide blood samples.

BRCA1 and BRCA2 mutation analyses are under way, led by Drs. Malone and Ostrander, for defined subsets of the case population, such as women diagnosed before the age 35 years and women with a first-degree family history of breast cancer (14,15). Because these analyses rely on a population-based ascertainment scheme, they avoid the somewhat limited generalizability associated with studies of high-risk women ascertained through genetic clinics or through families accumulated by referral because of their unusual and extreme profiles. In addition, the families of women with a mutation are being characterized to evaluate founder effects within these families. We have also been well positioned to investigate the contributions of a number of other candidate genes, including ATM and the ER gene.

Prior analyses of family history data from our first case–control study revealed that women with a first-degree family history of breast cancer had a 50% lower risk of dying of the disease than did women with no family history of breast cancer or women with a second-degree-only family history of breast cancer (16). As the number of breast cancer cases tested for BRCA1/BRCA2 increases, we have also begun to investigate the relationship of these factors to tumor characteristics and the risk of dying. This unique infrastructure built on a well-defined population-based series of cases and controls offers rich scientific benefits in a number of different directions.

**Methodologic Considerations**

**Framework for Developing Methodologies**

One immediate statistical benefit of having a design framework is that it allows a unified statistical approach to the analysis of complex data collected in multiple stages. The following discussion centers on the derivation of the likelihood functions that are useful for developing statistical methodologies.

**Likelihood for Data Collected at Stage A**

The data \((d_p, x_p)\), where \(d_p\) is an indicator for the proband disease status and \(x_p\) represents the corresponding covariates from probands, are generally informative about the associations between covariates and phenotype. Following usual statistical considerations, the likelihood function for retrospective data may be constructed as

\[
l(d_p, x_p) = \prod_i f(x_i|d_i),
\]

where \(f(x_i|d_i)\) denotes an individual contribution from the \(i\)th case or control proband. The above likelihood function may take a different form if stage \(A\) is a matched case–control study or a cohort study. The likelihood function for linkage-disequilibrium and loss of heterozygosity among cases can be similarly constructed, treating \(x_i\) as markers.

An important variation of the above likelihood is an extension that incorporates the family history information reported by probands. This additional information can be used to assess familial aggregation of phenotype, in conjunction with the association analysis. The likelihood function may be modified to

\[
l(D_r, d_p, x_p) = l(d_p, x_p)\prod_i f(D_r|x_r, d_i),
\]

where \(D_r\) represents a vector of phenotypes reported on relatives and represents a likelihood contribution from the family history as reported by a proband. The covariates of relatives can also be incorporated, but they may be of doubtful reliability when reported only by a proband. Note that the upper case letter is generally used here to denote a vector of variables from a family. Note that while the notation ‘\(i\)’ is used to denote the \(i\)th family, we also use letters ‘\(r\)’ and ‘\(p\)’ to denote relative and proband, respectively.

**Likelihood for Data Collected at Stage S**

Data collected from stage \(S\) can be used for either 1) association studies via a case-relative control study design; 2) linkage studies via affected sib-pairs; 3) combined association and aggregation analysis; or 4) combined association, segregation, and aggregation analysis. The likelihood functions to be constructed need to be specific to the objectives of the analyses. The likelihood contribution for the association analysis has been comprehensively dealt with by the Gene Characterization Panel and has been described by Witte et al. (11). For linkage analysis, the likelihood function is similar to those described below. Further, the likelihood function for the combined association and aggregation analysis is similar to the one indicated above after including covariates collected from relatives.

Here we describe a likelihood function for combined association, segregation, and aggregation analysis. Let \(g\) denote the putative gene that may contribute to familial aggregation. The objective is to estimate its penetrance function \(Prf(d|g,x)\) and its allele frequency. Under the typical assumptions for inherited factors, one can derive a joint distribution of putative genes as well as a joint distribution of phenotypes, covariates, and genotype. Then one may construct the following likelihood function,

\[
l(D_r, X_r, d_p, x_p) = \sum_{g_r G_r} l(D_r, X_r, G_r, d_p, x_p, g_r),
\]

where the summation is over all possible genotypes and the likelihood contribution \(l(D_r, X_r, G_r, d_p, x_p, g_r)\) on the right-hand side follows directly from the likelihood [2] by including putative genes, using analogous notation.

**Likelihood for Data Collected at Stage L**

Data to be collected at stage \(L\) are generally used for linkage or linkage-disequilibrium analyses. Let us consider the situation in which marker data are collected from only those relatives
identified at stage S. In this case, the likelihood function for the
data takes a very simple form and may be written as
\[ l(D_r, X_r, M_r, d_r, x_p, g_p) = \sum_{g_p, G_r} l(D_r, X_r, G_r, d_r, x_p, g_p) \]
\[ f(M_r, m_p, G_r, g_p), \]
where the second distribution on the right-hand side of the equa-
tion is specified by the linkage and linkage-disequilibrium pro-
cesses. Note that markers are assumed to be independent of
environmental factors given putative genes, an assumption that
could be violated in some cases in which markers are directly or
indirectly correlated with certain lifestyles.

There is an important modification to the above likelihood
function [4] to accommodate additional relatives ascertained at
stage L, who have not been part of stages A or S. Suppose that
an ad hoc procedure is used to ascertain these additional rela-
tives, i.e., there is no obvious way to quantify the ascertainment
rule. The additional data can be included in the above likelihood
function as
\[ \sum_{g_p, G_r, G_a} l(D_r, X_r, G_r, d_r, x_p, g_p) f(M_r, m_p, m_a | G_r, g_p, G_a) \]
\[ f(G_r | D_r, X_r, G_r, g_p, D_r, d_r), \]
where the third factor is contributed by the additional relatives.
When additional families are contributed from external sources
and are assumed to arise from the same conceptual population,
their genetic markers, along with disease phenotypes and cov-
ariates, can also be combined into the above likelihood func-
tion. Let the subscript “o” denote other families. The likelihood
function may then be constructed as
\[ \sum_{g_p, G_r, G_a} l(D_r, X_r, G_r, d_r, x_p, g_p) f(M_r, m_p, m_a | G_r, g_p, G_a) \]
\[ f(G_p | D_r, X_r, G_r, g_p, D_r, d_r) \sum_{G_o} f(M_o | G_o) f(G_o | D_r, X_o), \]
where the additional summation factor is contributed from those
external families. The key assumption made in the above like-
lihood function is that the estimated penetrance and allele fre-
cuency from the multistage study are applicable to those addi-
tional families. Under this assumption, additional families may
be sampled from high-risk family registries, from a genetic
counseling clinic, or from other ad hoc sources.

Sample Size and Power

At the planning stage, evaluating sample size is important to
establish the feasibility of a proposed study design. After a de-
sign with a feasible sample size is chosen, evaluating the study
power, after acknowledging the complexities of the disease of
interest, becomes critical. Although analytic power and sample
size formula would be desirable, the development of such ex-
pressions appears to be challenging because of the multiple
study stages and various analyses of interest.

Instead, we propose to use a Monte Carlo method to estimate
sample size and to evaluate power for studies with specific null
hypotheses and specific alternative hypotheses. After consider-
ing local resources and budgetary constraints, one may choose a
particular design that consists of design choices at stages A, S,
and L. Once a design is finalized, one can simulate data follow-
ning the proposed design protocol. Specifically, the simulation
may include 1) families with desired pedigree structures, 2) one
or more environmental factors, 3) one or more genetic factors
with known positions along the human genome, 4) genetic markers
with known density along the human genome, 5) phenotypes
that are determined by genetic and environmental factors, and 6)
any other ancillary features of proband identification, relative
ascertainment, missing data, and measurement errors. The simu-
lated data are analyzed using the methods alluded to above.
Repeating the above procedure for a large number (e.g., 1000) of
simulation runs allows one to estimate the power of key test
statistics for the proposed design. By profiling over a range of
sample size scenarios, one can then pick a sample size so that the
study will have the desired power. By following a similar pro-
cedure, one can evaluate the power of a study given the study
design and sample size.

Framework for Comparing Statistical Methodologies

Since the fundamental paper of Elston and Stewart (17),
many methods have been developed in statistical genetics. Most
of these methods are applicable to genetic analysis on data col-
lected from studies with integrated designs. Within this general
design framework, one can compare the relevant methods with
the likelihood methods outlined in the “Framework for Devel-
oping Methodologies” section.

It is anticipated that the likelihood methods outlined above
may turn out to be equivalent to, or similar to, some established
methods. In such cases, the general likelihood methods may be
thought of as an extension of established methods, in the sense
that existing methods are generalized to studies of other designs.
Conversely, established methods can be expected, sometimes, to
differ from these likelihood methods. In such cases, it is neces-
sary to compare them in term of bias, coverage probability, and
power (or efficiency). A favorable result for the likelihood meth-
ods helps their establishment. Unfavorable results are also help-
ful because they would identify areas for improving the likeli-
hood formulation. Theoretically, for designed studies, appro-
priately constructed likelihood methods should be able to
yield estimates of maximal efficiency.

Practicalities of the Likelihood Approach

As a general theoretical paradigm, likelihood has been a
foundation for almost all statistical methodologies (18). Without
exception, likelihood is the foundation underlying most of sta-
tistical genetics (17). Following the same principle, all of the
likelihood functions described above serve as an efficient tool
for synthesizing the statistical information contained in various
types of studies. In applications of the above likelihood func-
tions to actual studies, the likelihood approaches are expected to
face challenges, both statistically and computationally.

Statistical challenges are due primarily to the distributional
assumptions required by the likelihood approach. For example,
to construct a likelihood function (equation 3) for a segregation
analysis, one typically assumes conditional independence, i.e.,
familial aggregation of phenotypes is entirely accounted for by
the assumed putative genes. Consequently, the corresponding
likelihood function (equation 3) can be written as
\[ \sum_{g_p, G_r} l(D_r, X_r, G_r, d_r, x_p, g_p) \]
\[ = \sum_{g_p, G_r} \prod_{j \in \text{relatives}} f(d_j | x_r, g_j) f(x_p, g_p | d_p) \]
which has been explored in-depth separately (19). In studies of
monogenic diseases, this assumption of conditional independence
is justifiable and appropriate. The corresponding likeli-
hood methods have been useful in characterizing putative genes and in discovering those genes in past research. However, in studies of complex traits, this assumption becomes problematic, because there are many factors, other than the assumed putative gene, that may cause familial aggregation. The presence of such factors would violate this conditional independence assumption. Naively using the likelihood under conditional independence may lead to 1) biased estimation of segregation parameters (20) and 2) inflated type I error because of incorrect estimation of the standard errors (21–24). Overcoming this statistical challenge has motivated recent development of semiparametric methods (25–28).

Computationally, calculating the likelihood function, especially for large pedigrees and for a large number of marker loci, can be extremely challenging. The primary reason is that it requires enumeration over all possible alleles of putative genes on all family members in pedigrees, the computational burden of which could increase exponentially with the number of founders (29). However, multipoint linkage analysis with the use of the Elston and Stewart algorithm (17) needs to enumerate all possible phases at multiple marker loci, the computation burden of which could increase exponentially with the number of marker loci. Much of the past research effort aims at circumventing these computational difficulties. Among a large body of literature in this area, the Elston and Stewart algorithm (17) is the foundation for the traditional logarithm of the odds (LOD) score method. Simulation-based methods (30–33) have been developed. In the area of developing methods for multipoint linkage analysis in medium-sized pedigrees, Idury and Elston (34) have been quite successful. Even more recently, we have been developing a semiparametric method for multipoint linkage analysis (9,28). It seems evident that the further development of flexible, robust, and efficient data analysis tools will be an important element in an overall disease gene discovery and characterization program.

**DISCUSSION AND SUMMARY**

The anticipated completion of the entire Human Genome Map by year 2003 and ever improved high-throughput technologies, such as microarray techniques, will facilitate studies for discovering and characterizing disease genes in the postgenome era. In the beginning of this new era, it is timely to have this workshop on study designs. Indeed, this workshop provides an opportunity for us to present our thoughts on this important issue and has further allowed us to critically evaluate those study designs used in practice as well as those that have just been proposed in the workshop.

In this paper, we provide a rationale for integrated study designs and give examples of potential study designs within the framework of population-based family study designs. Furthermore, these designs are evaluated from genetic and epidemiologic perspectives as well as from a practical viewpoint. Designs introduced in this paper serve best as a stepping-stone for genetic epidemiologists to evaluate critically their study design options and to adopt and expand them as appropriate.

Although the above discussion focuses on genetic factors, the theme of the workshop, we should not overlook the important contribution of environmental factors to human diseases. Studying gene–environment interactions is also an important task for genetic epidemiology. When planning a study focusing on gene–environment interactions, one should consider not only those design issues described above but also epidemiologic issues. For example, environmental factors gathered on family members are likely subject to the influence of difference in ages, differences in birth cohorts, and differences in the periods of diagnoses, and thus some designs such as the case–family study may not be entirely appropriate for assessing gene–environment interactions.

Too often, study designs are advocated or dismissed too quickly. To ensure an accurate assessment of the study design, one needs to use statistical methods that require minimum nuisance assumptions and that are efficient for the intended analysis. Only then, with efficient methods, can one begin to assess the validity and the efficiency of the study design. Often, however, only suboptimal statistical methods are developed and are used to evaluate designs. In such cases, the conclusions regarding certain study designs could be misleading.

It is clear that developing efficient statistical methods is critical and is an integrated part of the development of genetic epidemiology. In the past few years, some effort has been made to develop methods for association analysis with correlated family data, for aggregation and segregation analysis using case–control family data and for linkage analysis incorporating environmental factors. How to combine these analyses, in the hope of gaining efficiency, remains an open question as does the incorporation of time-dependent outcomes and time-dependent exposure variables. As expected, developing these methodologies will be one of many active research areas, along with the general development of genetic epidemiology, in years to come.

**REFERENCES**


NOTES

1Editor’s note: SEER is a set of geographically defined, population-based, central cancer registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Registry data are submitted electronically without personal identifiers to the NCI on a biannual basis, and the NCI makes the data available to the public for scientific research.

Supported in part by Public Health Service grants CA93020, CA33619, CA53996 (National Cancer Institute); AG14358 (National Institute on Aging); GM28356 (National Institute of General Medical Science); and RR03655 (National Center for Research Resources), National Institutes of Health, Department of Health and Human Services.

We thank the participants in the workshop for their helpful comments and discussions.